A Probabilistic Approach To Non-Rigid Medical Image Registration

Ivor J. A. Simpson

St Edmund Hall

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2012

Abstract

Non-rigid image registration is an important tool for analysing morphometric differences in subjects with Alzheimer's disease from structural magnetic resonance images of the brain. This thesis describes a novel probabilistic approach to non-rigid registration of medical images, and explores the benefits of its use in this area of neuroimaging.

Many image registration approaches have been developed for neuroimaging. The vast majority suffer from two limitations: Firstly, the trade-off between image fidelity and regularisation requires selection. Secondly, only a point-estimate of the mapping between images is inferred, overlooking the presence of uncertainty in the estimation.

This thesis introduces a novel probabilistic non-rigid registration model and inference scheme. This framework allows the inference of the parameters that control the level of regularisation, and data fidelity in a data-driven fashion. To allow greater flexibility, this model is extended to allow the level of data fidelity to vary across space. A benefit of this approach, is that the registration can adapt to anatomical variability and other image acquisition differences.

A further advantage of the proposed registration framework is that it provides an estimate of the distribution of probable transformations. Additional novel contributions of this thesis include two proposals for exploiting the estimated registration uncertainty. The first of these estimates a local image smoothing filter, which is based on the registration uncertainty. The second approach incorporates the distribution of transformations into an ensemble learning scheme for statistical prediction. These techniques are integrated into standard frameworks for morphometric analysis, and are demonstrated to improve the ability to distinguish subjects with Alzheimer's disease from healthy controls.

Acknowledgements

This thesis would never have been completed without the help and support of several key people. In particular, I would like to thank:

Firstly, my supervisors Julia Schnabel and Mark Woolrich, who both provided guidance, encouragement and support when I needed it. They patiently listened to all my half-baked ideas, and helped me to focus on the best of them. I am grateful to Julia for explaining the world of registration to me, setting me goals and deadlines and her willingness to edit conference submissions, even at ridiculous hours. I want to thank Mark for teaching me about probabilistic modelling, why Bayes' is best and for helping me to take a step back from the details, and think about what I'm actually trying to achieve.

Adrian Groves for helping me through every step of this DPhil, patiently explaining the answer to every question I had and giving up a huge amount of time to discuss all of my ideas in detail. Without Adrian, this thesis would have been much less interesting and taken a lot longer.

Jesper Andersson for his kind assistance and providing an excellent codebase in FNIRT, which formed the basis for the implementations in this thesis.

The FMRIB IT guys, Dave Flitney and Duncan Mortimer, for maintaining a fantastic cluster, and their continual patience with me breaking it.

ADNI for making such a vast collection of imaging data publicly available.

Everyone involved in the Life Sciences Interface Doctoral Training Centre, for funding my DPhil and giving me the opportunity to learn more about science.

The Guarantors of Brain, St Edmund Hall, the DTC, the Department of Engineering and Julia Schnabel for helping me with travel funding, so the work in this thesis could be presented at conferences.

The students and post-docs at the IBME for their comradeship, banter and discussions. In particular, my MICCAI road-trip buddies, Mattias Heinrich and Manav Bhushan, for so many useful conversations about image registration.

All the DTC gang who've stuck together for the last four years, providing endless quantities of entertainment and moral support.

My friends from Teddy Hall MCR, for showing me that a DPhil at Oxford is not just about research, there are posh dinners and croquet games too.

My Mum for inspiring me to do something with my life and constantly encouraging me to achieve, even when I didn't want to.

Finally, and most importantly, my wife Laurie for her endless patience, reassurance and support. I couldn't have done it without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the recent availability of large, high resolution medical image datasets, the manual analysis of every image by trained experts is becoming intractable. To deal with this plethora of data, automated medical image analysis techniques have been devised. These tools are designed to extract application specific details from the data, either automatically, or with limited manual intervention. The development of such tools is essential for accurate information to be obtained, and compared across subjects to evaluate the characteristics of a disease in a population. Once extracted, the image derived information can be used as a basis for objective diagnosis or prognosis of individual patient outcomes.

This thesis considers the analysis of differences, and changes over time, in the morphology of the human brain when affected by Alzheimer's disease (AD). To find robust and sensitive morphological imaging biomarkers of neurodegenerative disease, the differences and similarities between and across population groups need to be statistically analysed. In order for such an analysis to be performed, all of the image data need to be placed within a common frame of reference. Once the images are transformed to a common reference frame, a particular coordinate location has a common meaning across the set of images [177]. In neuroimaging group studies, subject images are usually transformed to an average population image space, often referred to as an atlas. There are several standard atlases that are used in neuroimaging, one of the more common is the MNI152 atlas [54]. More task specific atlases can also be derived, which is discussed in section 2.3.3.

The process of estimating the mapping from a subject image to an atlas space is referred to as spatial normalisation. Spatial normalisation reduces inter-subject anatomical variability and enables meaningful comparison of image information [58]. It also provides a framework for analysing morphometric deviation

between subjects and an atlas [11][38]. Spatial normalisation is a challenging problem, and is addressed through the use of image registration algorithms.

The development of advanced tools for registration of structural magnetic resonance images (MRI) of the brain, is the main focus of this thesis. The next section introduces medical image registration and describes some of the limitations of current approaches. This is followed by a brief introduction to MRI, and the benefits of image analysis in AD.

1.1 Medical Image Registration

Medical image registration is the problem of estimating anatomical, or functional correspondence between images. This correspondence takes the form of a mathematical mapping between the two image coordinate systems. Image registration has a broad variety of medical applications [84], and has been implemented using a wide range of methodologies (see [121][207] for reviews).

There are two particularly important roles that image registration plays in neuroimaging: Firstly, it allows the spatial normalisation of multiple subjects into a common reference frame. The estimation of this mapping is sometimes referred to as inter-subject registration. Secondly, registration can be used to align images of the same subject, known as intra-subject registration. These images may have been acquired at different times, or using different imaging modalities. As such, intra-subject registration provides a mechanism for describing anatomical changes in a subject over time. Both inter-subject and intra-subject registration are explored in this thesis.

Brain image registration methods are typically driven by the alignment of image intensities, although some methods have been developed that are based on specific landmarks [26][134]. However, the determination of a homologous, or sufficiently descriptive, set of landmarks to relate two images is usually only semi-automated, and many landmarks may be required for a reasonable mapping. An alternative approach, is cortical curvature matching [45][55]. These methods estimate the deformation of the cortical surface with the aim of aligning regions that have a similar surface curvature. However, such approaches do not seek to align subcortical structures, such as the hippocampus, which is strongly associated with AD [97]. Therefore, this thesis only considers the class of image intensity based registration methods. Image intensity based registration approaches optimise a chosen transformation model such that the anatomical or

functional correspondence, as measured from the voxel intensities, is maximised subject to any constraints.

1.1.1 Global Registration Methods

Early approaches to image intensity based registration in neuroimaging were used to correct for global differences, such as position, scale and shearing [192][39]. The choices of transformations vary from rigid transformations with six degrees of freedom (rotations and translations in three directions), to affine transformation with up to an additional 6 parameters (scales and skews in 3 directions). These global registration methods are able to account for the gross anatomical differences between images, and provide a coarse level of alignment. Global registration methods have been well studied, and many efficient implementations are available [15][193][98]. However, in order to provide a high-resolution spatial normalisation, or model morphometric differences between images, a more flexible class of methods is required.

1.1.2 Non-Rigid Registration

Intensity driven non-rigid (sometimes referred to as non-linear) registration was introduced to neuroimaging to resolve spatially local changes in brain shape [63]. These methods have far greater degrees of freedom than global registration, and as such allow for a more flexible mapping between images. A global registration is often a required pre-processing step for non-rigid registration. This is because the reduced degrees of freedom make it less susceptible to sub-optimal solutions.

1.1.3 Issues In Non-Rigid Registration

Intensity based image registration methods measure correspondence according to the differences of the image intensities at each location. An approach that purely minimises intensity differences, will probably infer large, complicated, and noisy transformations. This is because voxel intensities are a noisy measurement of the tissue properties of the underlying anatomy. As the same tissue types are found across the brain, the voxel intensities do not provide sufficient information for a distinct anatomical correspondence. Moreover, small differences in intensity, which could result from acquisition noise, or image processing such as subsampling, or smoothing of the image data, will be attempted to be matched.

This is a key issue when using a more flexible transformation, as the model can more closely match small image differences. An unconstrained approach to non-rigid registration may result in a wholly inaccurate registration, attempting to match non-corresponding structures despite them being geometrically different, and anatomically distinct, because of irrelevant intensity differences. Therefore, to infer an accurate and plausible transformation, some prior knowledge needs to be introduced as a means of regularisation. An issue present in all approaches to non-rigid registration, is how to balance the importance of matching image intensities, and the strength of the prior knowledge. Methods to address this issue are addressed in chapters 3 and 4, with results given in chapter 5.

A further concern is that intensities necessarily do not specify a unique optimal mapping between images, for any given transformation model. The ambiguities in voxel intensity matching can be calculated, and incorporated into a quantification of the uncertainty in the inferred mapping between images. Understanding the uncertainty in registration is an important consideration in any statistical analysis where registration is required. This is explored in chapters 6, 7 and 8.

Both of these problems can be addressed through the use of a probabilistic model for image registration. Therefore, chapter 2 includes a brief review of the components required for a non-rigid registration algorithm, with a focus on the construction of probabilistic models of registration.

The next section of this chapter introduces the context for the application of medical image registration in this thesis.

1.2 Alzheimer's Disease

AD is a progressive neurodegenerative disease, and the most common to be associated with the symptoms of dementia. Individuals affected by AD suffer a progressive impairment of cognitive function, and emotional disturbances due to the accelerated rate of death and degradation of neurons and synapses. AD is a degenerative and terminal condition linked to ageing. AD currently affects an estimated 26.6 million people worldwide [31]; consequently AD is responsible for huge medical costs and poses a considerable social burden. There currently exists no cure for Alzheimer's disease. This may be partially due to the imprecise tools that are commonly used as a surrogate measure for the effects of any developed treatments.

AD has been shown to be associated with the accumulation of abnormal proteins in the brain, which is accompanied by progressive synaptic, neuronal and

axonal damage. These changes may begin to occur several years prior to any symptomatic effects. The presence of abnormal proteins and grey matter atrophy is initially observed in the medial temporal and parietal lobes. Subsequently, the frontal lobe is affected in the later stages of the disease [159].

AD is often preceded by the patient suffering from mild cognitive impairment (MCI). Patients with MCI often suffer from mild memory loss, but are still capable of everyday living [136]. MCI may be caused by several different factors. However, patients with MCI have an increased likelihood of progressing to AD, at a rate of 10-15% per year [118]. For this reason, patients with MCI are often studied when investigating the early effects of AD.

Diagnosis of AD is often based on insensitive symptomatic measures such as the clinical dementia rating [128]. Such tests are poorly suited for detecting changes in mildly symptomatic subjects, are insensitive to small changes in behaviour, and may not be suitable for ascertaining which specific pathology has lead to an incidence of dementia. Tests on the cerebrospinal fluid (CSF) for AD have been developed [6]. These tests have been shown to be sensitive to AD. However, it is still unclear how specific these tests are [46]. Obtaining samples of the CSF for these tests requires an invasive procedure, making this unsuitable for widespread use.

Radiological imaging of subjects suffering the symptoms of dementia can be used to provide a differential diagnosis. Furthermore, imaging can potentially provide measures that are better suited for measuring the efficacy of a treatment than cognitive, or functional scales.

1.2.1 Imaging in Alzheimer's Disease

Radiological imaging of patients with suspected AD enables an observation of the current state of brain function or anatomy. The information provided by imaging can yield objective evidence regarding the state of pathology that would otherwise be unavailable. Quantifiable features derived from imaging that are indicative of biological state, are known as image derived biological markers, or more simply as imaging biomarkers. The International Working Group for New Research Criteria for the Diagnosis of AD [50] proposed the use of imaging biomarkers as supporting evidence for the diagnosis of AD. Imaging biomarkers can also be used as a measure of treatment efficacy [96]. Currently, there are no widely accepted imaging biomarkers of disease progression [61], although preliminary

studies suggest that such an approach could provide greater statistical power than traditional measures [89].

The selection of a suitable set of biomarkers requires some consideration. A useful biomarker needs to be both biologically plausible, and sensitive to changes in disease progress. As the pattern of change varies over the course of the disease, an ideal set of biomarkers would reflect this.

Several imaging modalities have been previously used in the diagnosis of AD. These include functional imaging techniques such as positron emission tomography (PET) using an FDG tracer. FDG PET allows the visualisation of glucose uptake, and therefore provides a measure of localised brain activity [82]. Functional magnetic resonance imaging (fMRI), which measures the blood oxygen level dependence (BOLD) signal, has also been used [153]. Molecular imaging approaches such as PET using the PIB tracer have been used to image the amyloid proteins associated with AD [107]. Micro-structural imaging approaches, such as diffusion tensor magnetic resonance imaging (DTMRI), have been used to visualise the integrity of neuronal connections in AD [154].

Structural magnetic resonance imaging (MRI) allows a view of the anatomical macro-structure, and provides a means for assessing the shape and size of brain structures. This thesis focuses on the use of structural imaging methods, which has the advantage of the availability of large public datasets. There is a potential for AD biomarkers derived from other imaging modalities to be used in combination with structural MRI information, as they will contain complementary information. The next section provides some details regarding the acquisition of MR images, before elucidating on their uses in AD.

1.2.2 Magnetic Resonance Imaging

Magnetic resonance imaging is a powerful non-invasive method of medical imaging, which in contrast to some other structural imaging methods such as computed tomography does not use ionizing radiation. This makes MRI ideal for both clinical and scientific work. MRI also provides a high contrast in soft tissue anatomical structures, which is appropriate for imaging the human brain.

MRI scanners use a strong magnetic field, most commonly between 1.5-3 Teslas for neuroimaging. This aligns the magnetic spin of hydrogen nuclei of a body placed within the field. Radio frequency (RF) pulses, produced at the resonance frequency of the target protons, are then applied to "flip" the angle of magnetic spin. These flipped protons create a magnetic field perpendicular to the main

field. This induces currents in a set of receive coils, allowing the measurement of the perpendicular component of the proton's spin. The "relaxation" of these hydrogen nuclei to their original magnetisation (aligned with the field) is then measured according to two time periods. T1, the time taken for relaxation of the magnetisation in the direction of the magnetic field, and T2 the time taken for the loss of the magnetisation perpendicular to the main magnetic field. The MR image is constructed from the estimated measurements of these time periods at each point in space. The relaxation time for these two factors varies between different tissue types, and it is this variation that provides the contrast between tissue types in the MR images [80]. Image intensity or contrast may vary across the image in a spatially smooth manner according to an unknown intensity bias field [188]. Amongst other things, intensity bias fields may be caused by inhomogeneous sensitivity of the receive coils to locations in the brain. MR produces images that are corrupted by Rician noise, which for typical values of SNR is approximately Gaussian [74].

In order to create a 3-D image, positional information needs to be encoded. This is achieved by applying an additional magnetic field gradient across space. The frequency that a proton "precesses" or spins at, is proportional to the strength of the magnetic field. This is known as the Larmor frequency. The use of a magnetic field gradient causes protons in different locations to spin at different frequencies. By exciting, and then "listening" to these different frequencies, the T1 and T2 time periods can be acquired across different locations.

MRI uses different pulse sequences in order to acquire images with varying contrasts. The choice of the sequence is dependent on the application. In neuroimaging, it is common for T1-weighted structural images to be used. However, T2 sequences also have their uses and provide good contrast for some tasks, for example in locating lesions. The resolution of the acquired MR images is dependent on the sequence, the scanner field strength and the required signal-to-noise ratio. A typical image resolution from a 3T scanner is 1 mm^3 per volume element (voxel).

1.2.3 Structural MRI in Alzheimer's Disease

Structural T1-weighted MRI of the brain has been commonly used as a modality to develop imaging biomarkers of AD [61]. MRI allows the measurement of macroscopic changes in brain structure. This can be used for the quantification of both global [59][172], and local [60][159][89][57], changes in brain morphology

(commonly atrophy). It can also be used to describe morphological differences between subjects [11][106][120][44][90][179]. Changes in brain morphology can include differences in size, and shape of brain structures. By analysing the patterns of morphological differences across several subjects, a consistent set of biomarkers that differentiate AD from normal ageing, or MCI, can be found. Morphological biomarkers associated with AD have been established as being visible from MRI prior to the observation of clinical symptoms [142], and hippocampal atrophy has been shown to correlate well with CSF markers [76].

As MR imaging biomarkers of AD are visible before symptomatic effects, these could be used to facilitate an early diagnosis, or prognosis. This could also provide an insight into the biological mechanisms underlying AD. Perhaps more importantly, imaging biomarkers can be be used as a quantitative measure of the efficacy of any disease modifying drugs on the underlying anatomy [96][56]. The predictive capabilities of structural MR images of the brain for the detection of AD are explored in chapters 7 and 8.

1.3 Summary Of Remaining Chapters

This thesis introduces two novel probabilistic models and inference strategies for non-rigid registration, these provide a solution to some of the limitations of current methods. The benefits of the proposed probabilistic registration frameworks are evaluated in terms of deriving robust spatially normalised morphological features that are predictive of AD.

Chapter 2 introduces the required components of non-rigid registration, focusing on how a probabilistic registration model can be formulated, and how this fits within the existing registration literature. Bayesian inference strategies on probabilistic models are briefly reviewed. The chapter concludes with a discussion of how morphological features can be extracted from MR images of the brain using non-rigid registration, and how these features can be statistically analysed.

Chapter 3 presents a novel probabilistic model and inference strategy, for non-rigid registration. Bayesian inference of this model is described, which allows the level of regularisation, and data fidelity, to be inferred from the data. This eliminates the need for manual tuning of the model parameters by allowing the registration to adapt to the presented data. Furthermore, this model intrinsically provides estimates of the uncertainty of the registration, this is investigated further in chapter 6.

Chapter 4 further develops the probabilistic non-rigid registration model that was proposed in the previous chapter, and adapts it to describe spatially varying estimates of model mismatch. This allows the data driven inference of a spatially varying trade-off between data fidelity and regularisation.

Chapter 5 introduces methods for the quantitative validation of non-rigid registration, and provides results for the two proposed registration framework in comparison to a pre-existing tool. The probabilistic registration frameworks are shown to adapt to image acquisition, and anatomical, differences and are demonstrated to provide robust and accurate inter-subject registration.

Chapter 6 discusses the concept of registration uncertainty, its interpretation, and how it can usefully be applied to resolve residual mis-registration in inter-subject registration. A novel adaptive smoothing filter, which is derived from the estimated registration uncertainty, is introduced. The use of smoothing to compensate for mis-registration is motivated in terms of providing an improved trade-off between sensitivity and specificity, with respect to propagated anatomical segmentation accuracy. The adaptive smoothing filter is shown to outperform several isotropic Gaussian smoothing kernels.

Chapter 7 investigates the application of the probabilistic non-rigid registration framework for the longitudinal image analysis of subjects with AD. Maps of localised expansion and contraction are estimated. These are shown to provide information that accurately discriminates between AD and age matched healthy controls. Furthermore, the spatial normalisation of these features was performed using the proposed registration framework. This permitted the adaptive smoothing filter, as described in chapter 6, to be used to compensate for mis-registration. The application of this filter is demonstrated to provide the most accurate classification of disease status from the spatially normalised longitudinal features, as compared with a range of isotropic Gaussian smoothing kernels.

Chapter 8 proposes the use of registration uncertainty as a novel mechanism for generating variability in statistical predictors. These predictors are amalgamated into an ensemble learning framework, which compensates for registration uncertainty. This framework is applied to the classification of subjects with AD from age matched controls based on two forms of estimated morphological brain differences. The variability in statistical predictors, as generated from samples of probable registrations, is demonstrated to improve classification, and provides better estimates of the uncertainty in prediction than traditional approaches.

Chapter 9 summarises the contributions of this thesis, discusses some ideas for future work and draws some overall conclusions.

Chapter 2

Background

This chapter begins by describing how registration can be formulated as a generative model, which allows for a probabilistic description of the model parameters. This is followed by an introduction of the constituent parts of a non-rigid registration algorithm, with a focus on components that fit within a generative model formulation. Approaches to Bayesian probabilistic inference are then briefly reviewed, and an appropriate inference framework for the problem of non-rigid registration is introduced. The chapter concludes with a discussion of how morphological features can be extracted from MR images of the brain using non-rigid registration, and how these features can be statistically analysed to predict a subjects' disease status.

2.1 Generative Models for Image Registration

Generative models provide a description of what the observable data should look like, given a set of parameter values. Such models usually incorporate a random noise model, which probabilistically describes the level of mismatch. Mismatch between the observed data, and the generative model of that data, is referred to as residual or error. For a complex problem, like inter-subject brain registration using real data with limited signal-to-noise ratio, residual errors are to be expected. A generative model allows these errors to be described explicitly. The reliability of a model parametrisation can thus be assessed through the use of a random noise model. As such, generative models provide a probabilistic approach to modelling, which facilitates model parameters being expressed as random variables that follow a probability distribution, rather than point estimates.

Image registration can, and has, been formulated as a generative model, but this is not necessarily the case, and many registration approaches seek to optimise the transformation parameters according to some data derived criteria. The following sections describe the requisite components of a registration method, and explain how a generative formulation of registration can be achieved, including any limitations or benefits of such an approach.

2.1.1 Transformation Models

A key feature of any non-rigid registration algorithm is the choice of the transformation model that describes the mapping between images. Many transformation methods have been proposed for use in the field of medical image registration, for which a broad review can be found in [86].

Consider the problem of registration between a moving source image, $\mathbf{x} : \Omega_x \rightarrow \mathbb{R}$ to a fixed reference image, $\mathbf{y} : \Omega_y \rightarrow \mathbb{R}$, where $\Omega_x, \Omega_y \subset \mathbb{R}^3$ are the domains of the 3-dimensional source and target image, respectively. Registration estimates the transformation that projects $\mathbf{x}$ onto the domain of Ω_y. This is achieved by inferring the mapping from the reference image domain, Ω_y, to the source image domain, Ω_x. This mapping takes the form of a deformation field $\mathbf{u}$:

$$\mathbf{u}(\mathbf{c}, \mathbf{w}) : \mathbf{c} \in \Omega_y \times \mathbf{w} \in \mathbb{R}^{N_t} \rightarrow \mathbf{c}_x \in \Omega_x \tag{2.1}$$

where $\mathbf{u}(\mathbf{c}, \mathbf{w})$ encodes the deformation field as a change in coordinates that is defined at each point, $\mathbf{c}$, in the reference image domain, Ω_y. The deformation field is dictated by a set of N_t transformation parameters, $\mathbf{w}$, and results in the corresponding co-ordinate $\mathbf{c}_x$ in the source image domain. The relationship between $\mathbf{w}$ and $\mathbf{c}_x$ is described by the choice of transformation model. Using the same notation, transforming $\mathbf{x}$ to Ω_y can be written as:

$$\mathbf{t}(\mathbf{x}, \mathbf{w}, \mathbf{c}) = \mathbf{x} \circ \mathbf{u}(\mathbf{c}, \mathbf{w}) \rightarrow \mathbb{R} \tag{2.2}$$

where $\circ$ refers the composition operatior in this equation. This function takes the source image, $\mathbf{x}$, transformation parameters, $\mathbf{w}$, and a coordinate in the reference domain $\mathbf{c}$, and returns the intensity of the corresponding location in the source image $\mathbf{c}_x$. In all further equations, $\mathbf{c}$ is omitted and $\mathbf{t}(\mathbf{x}, \mathbf{w})$ refers to the complete transformed source image, for all points in the domain Ω_y.

$\mathbf{t}(\mathbf{x},\mathbf{w})$ can be used as a forward model in a generative model of registration as it generates the model prediction of the observed target image, based on the estimated transformation of the source image. This formulation is not restricted to use in generative models, and is used in many approaches to registration.

To provide estimates of uncertainty of the transformation parameters, $\mathbf{w}$, the correlation of their effects on the transformed image, or cost-function is required to describe their joint distribution. As registration uncertainty is one of the issues with which this thesis is concerned, transformation models that allow efficient methods of calculating the correlation in $\mathbf{w}$ are preferable.

Parametric Transformation Models

A broad class of transformations suitable for use in non-rigid registration are "parametric" transformation models. Parametric transformation models describe the complete image transformation using a linear combination of basis functions [127]. The selected basis set can have either global [10], or local [156] support within the image. These methods fully describe the transformation through a single deformation field, with three directional components.

Parametric transformation models only guarantee a smooth mapping for small deformations, hence these are referred to as small deformation models [124]. When modelling larger transformations using a small deformation model, it is possible for the mapping to be non-smooth at points. This permits image "folding", points where the determinant of the deformation field Jacobian matrix, the 3×3 matrix of first order deformation derivatives, is less than or equal to zero. The capability of parametric models to resolve larger transformations can be improved through the use of hierarchical registration schemes [160]. Hierarchical approaches to registration are described in 2.1.5.

The primary advantage of using a basis set is that $\mathbf{w}$ is a compact description of the transformation. This results in a reduced set of parameters to estimate. These approaches often provide the facility for rapid calculation of the correlation between parameters as shown for the discrete cosine transform [10], or cubic B-splines free-form deformations [5].

B-spline Free-Form Deformations

The use of a B-spline basis set is preferable to a discrete cosine basis set, as the local influence of the B-spline simplifies the calculation of the correlation of $\mathbf{w}$ in high-resolution registration. A free-form deformation (FFD) transformation

model based on B-splines gives rise to an interpretation of each parameter in $\mathbf{w}$. Free-form deformations are described by the movement of control points, which for image registration, are normally arranged on a uniform grid across the image. Each control point can move in 3 dimensions, and the resulting deformation is calculated by blending the displacements of the control points using B-splines, which gives a smooth deformation across space [109]. $\mathbf{w}$ describes the movement of the control points, with a separate parameter for each direction. This is equivalent to stating that $\mathbf{w}$ contains the coefficients of a B-spline basis set. B-splines have a further advantage in being analytically differentiable, up to a degree based on the order of the B-spline. Although they are smoother, higher orders of B-splines give each control point a wider influence on the deformation across space, which complicates calculation and may lead to overly smooth deformations. Cubic B-splines are the most common to be used in registration [5][156], and produce a smooth and C^2 continuous deformation field, meaning that the second derivative of the transformation is continuous.

Models Using Multiple Transformations

There are several approaches to registration that estimate, and apply several different deformation fields to find the mapping from Ω_y to Ω_x, for example viscous fluid models [37], demons [180] and diffeomorphic demons [186]. Such approaches allow great flexibility in terms of the estimated transformation. However, the final mapping may be difficult to constrain in a principled manner, and it may be almost impossible to obtain meaningful estimates of the correlation between transformation parameters. This is because each deformation field will be defined using a separate $\mathbf{w}$, which will typically be a vector with between 10,000 and 2,000,000 components. This could lead to many million degrees of transformation freedom in total. In such a case, the joint distribution between the transformation parameters will be prohibitively difficult to calculate, and the complete transformation will be difficult to constrain.

Vector Field Approaches

In recent years, vector field approaches to medical image registration have become very popular. These allow a diffeomorphic mapping between images. A diffeomorphic mapping is one where the function $\mathbf{u}(\mathbf{c}, \mathbf{w})$, which maps from Ω_y to Ω_x, is bijective and smooth, and has a smooth inverse. The large deformation diffeomorphic metric mapping (LDDMM) [22] has been particularly influential.

In LDDMM, a vector field is used to define the transformation, and the complete mapping can be calculated by integrating the vector field over time. The LDDMM formulation allows the vector field to vary with time, thus very complex warps can be calculated. The advantage of such an approach is that by integrating over sufficiently small time steps, the transformation can be guaranteed to be diffeomorphic. However, in the standard LDDMM formulation a very large number of parameters are needed, as a three dimensional vector field for each time step is required.

A simplification of LDDMM was proposed by Ashburner in the tool diffeomorphic anatomical registration using exponentiated lie algebra (DARTEL) [7]. This approach uses a velocity field that is stationary in time. Although such an approach is less flexible than LDDMM, it still allows large deformations that are diffeomorphic and requires far fewer parameters. More recently, velocities that are variable in time have been parametrised as an initial momentum field using the principle of geodesic shooting [14]. Both of these approaches have implementations that allow fast computation of the correlation in $\mathbf{w}$. Although these approaches have fewer degrees of freedom than LDDMM, there are still far more parameters to estimate than in an approach using a parametric basis set.

Symmetric and Groupwise Registration Formulations

Registration can alternatively be formulated as a symmetric [8][150][17][111], or more generally as a groupwise [23][100], problem. Symmetric registration methods are unbiased in the choice of which image is selected to be the source, or the target, producing an identical output regardless. Groupwise registration aims to estimate the mappings from each subject to an intermediate image, which is unbiased to any of the original images.

Symmetry places an inverse consistency constraint on the registration [36], which means that the transformation between image $\mathbf{A}$ and image $\mathbf{B}$ should be the inverse of that from image $\mathbf{B}$ to image $\mathbf{A}$. This implicitly assumes a diffeomorphic mapping between images. Symmetric approaches should aim to infer transformations where the forward and inverse are equally probable [8], which requires an appropriately re-formulated cost function [17]. Although most approaches to symmetric registration require an inverse transformation, equivalent perturbation methods [111][176] do not, although Reuter et al. [140] caution that interpolation asymmetry may cause bias, particularly in longitudinal imaging, due to the additional smoothness introduced by image warping. Symmetry has been approached by registering both images to a midpoint between images [21][17][139],

thus treating them equivalently. Symmetric approaches to registration clearly provide a laudable constraint and the potential for extending this work to a symmetric formulation is discussed in chapter 9.

Groupwise is the generalisation of a symmetric registration, where multiple images are co-registered to an unbiased average space. Several constraints and approaches have been proposed [23][100][2][157]. Careful implementation is required, as these methods may require large amounts of memory and computational power.

2.1.2 Cost Function

The process of intensity based non-rigid registration can be thought of as the minimisation of a general cost function C:

$$C(\mathbf{w}, \phi, \lambda) = -\phi C_{sim}(\mathbf{y}, \mathbf{t}(\mathbf{x}, \mathbf{w})) + \lambda C_{reg}(\mathbf{w}) \tag{2.3}$$

where C_{sim} measures the similarity between two images and C_{reg} provides a regularisation cost, penalising unreasonable mappings. C_{reg} measures the deviation between the current transformation parameters, $\mathbf{w}$, and any prior knowledge that has been specified regarding the values of $\mathbf{w}$. ϕ models the fidelity of the image data, indicating how much it should be trusted. λ controls the strength of the regularisation. The trade-off between regularisation and image fidelity is very important, and is discussed in detail in the following section. In some implementations, this trade-off is expressed using a single parameter that models their ratio. The next section describes some typical types of regularisation that are used in non-rigid registration.

2.1.3 Regularisation Constraints

Registration is driven by an image similarity function that attempts to find the mapping that maximises the overlap of image information between two images. In structural MR images, the image intensities provides a noisy surrogate for anatomical tissue type. Simply finding the maximum value of the similarity function provides no guarantee as to the reasonableness of the inferred mapping, as voxel intensities do not provide sufficient information for plausible biological correspondence. Non-rigid registration mappings can be arbitrarily complex, and a close to perfect match of image intensities could be achieved that is biologically

meaningless. As a trivial example, the completely useless registration tool [151] maximises the similarity measure by transforming each voxel in the source image, to the position of the voxel in the target image that has the same rank when sorted by intensity. This implausible and unconstrained mapping between anatomies is clearly useless, despite maximising the image similarities. It is clear that to ensure the inference of a reasonable mapping, some prior information is required to constrain $\mathbf{w}$ to plausible values.

The trade-off between data-fidelity, ϕ, and regularisation strength, λ, is a key issue in all approaches to non-rigid registration. The λ parameter indicates the expected complexity of the inferred transformation, as measured by the form of the regularisation. As such, the trade-off between ϕ and λ describes the expected level of transformation complexity, for a given measurement of image similarity. If the relative importance of λ is too low, the optimisation may become under-constrained and large changes in transformation may be made for a small improvement in image similarity. This is because a more complex transformation was expected a priori than was actually required to adequately describe the registration. Conversely, if the regularisation strength is too high, the inferred mapping is likely to be overly smooth, and inaccurate. A motivating example of the effects of regularisation is given in Figure 2.1.

It is a necessary, but not sufficient condition that the transformation is spatially smooth to maintain the topology of the original image after transformation. The preservation of topology encourages spatially adjacent features in the original image to remain adjacent in the transformed image. It is also appropriate to penalise the complexity of a registration to ensure the plausibility of a mapping. This approach of penalising the path length, or deviation from the identity transformation of the inferred mapping, is used in several recent diffeomorphic works on registration [22][7][17][14]). It is clear that the smoothest, shortest mapping, which leads to an equivalent model fit, is preferable. When using a small deformation framework, regularisation is also used to reduce any folding of the image that may occur in complex or noisy transformations. An appropriate prior on the transformation parameters is one that encourages the transformation to be small and smooth.

Regularisation Priors

As the deformation field is wholly defined by the transformation parameters, $\mathbf{w}$, the regularisation of these parameters regularises the transformation. A prior on the distribution of transformations parameters can be specified as a multivariate

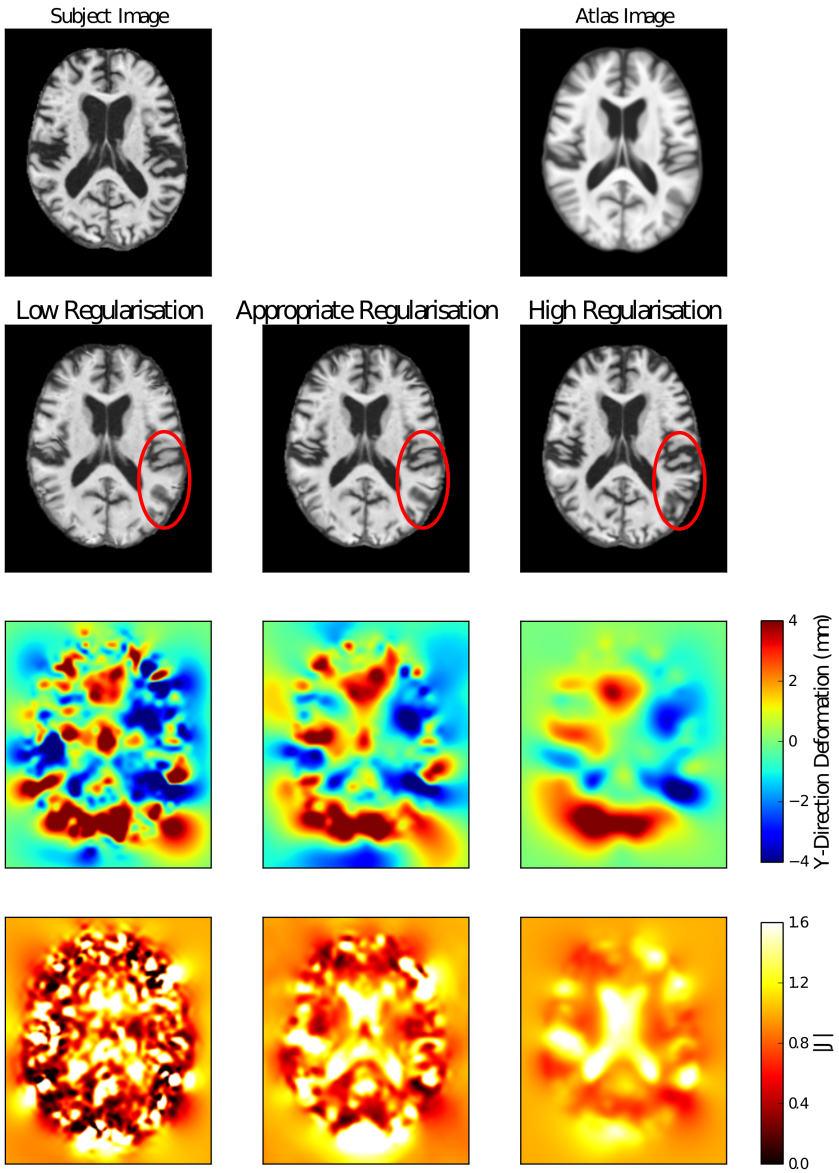

Figure 2.1: An example of inter-subject registration between a subject with Alzheimer's disease (top-left), to a representative atlas (top-right) under different levels of regularisation. The transformed images are shown in the second row. The example with a high level of regularisation looks quite different from the atlas, unlike the other two images. The deformation fields for the Y-direction are given in the third row. The Y-direction is chosen as it shows the greatest distinction between the registration methods. With a high level of regularisation, the transformation is very smooth, whereas for a low level of regularisation, it is very noisy. This results in a high level of image folding as shown in the determinant of the Jacobian map on the bottom row. Conversely, an appropriate level of regularisation leads to no folding in this case.

normal distribution, $\mathcal{N}$, which is parametrised by a mean and a covariance matrix, $P(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, (\lambda\Lambda)^{-1})$, where Λ is the inverse covariance matrix scaled by λ.

To penalise against the magnitude of the transformation, the mean of the prior distribution can be set as the identity transformation. The diagonal of the covariance matrix describes the expected variance of the transformation parameters from the identity. To support a smooth transformation, an appropriate covariance structure for the prior on $\mathbf{w}$ must be selected. The covariance terms of the prior encourage parameters in $\mathbf{w}$ that control spatially nearby regions of the deformation field, to take similar values, which in turn results in a smoother transformation.

Several forms of regularisation on the deformation field can be encoded into matrices for use as a prior on $\mathbf{w}$. These priors are encoded as spatial covariance, or inverse covariance matrices. The inverse covariance matrix can be thought of as a difference matrix, allowing differential operators to be encoded in an interpretable and sparse form. Conversely, the covariance matrix may be full, which in high resolution registration would be computationally intractable to store. Once appropriately encoded as an inverse covariance matrix Λ, the regularisation energy of the transformation, can be calculated using linear algebra as: $\mathbf{w}^T\Lambda\mathbf{w}$, where $\mathbf{w}$ is a vector containing the transformation parameters. Regularisation models can be efficiently applied though convolvution of the deformation field with the Green's function of the linear differential operator [30].

In the absence of any biological priors on the true covariance structure of the transformation parameters, there are a variety of structures that have previously been demonstrated as priors in non-rigid registration. These include membrane [4], thin-plate spline [26] and linear elastic [125] energy. These forms of regularisation have been adopted in many registration frameworks including [156][10][7][5]. The particular model of interest in this thesis is the thin-plate spline bending energy, which is defined for a deformation field, $\mathbf{u}$, in a direction, d, as follows:

$$\int_0^X \int_0^Y \int_0^Z \sum_{d=1}^{3} \lambda \left\{ \left(\frac{\partial^2 \mathbf{u}_d}{\partial x^2} \right)^2 + \left(\frac{\partial^2 \mathbf{u}_d}{\partial y^2} \right)^2 + \left(\frac{\partial^2 \mathbf{u}_d}{\partial z^2} \right)^2 + 2 \left[\left(\frac{\partial^2 \mathbf{u}_d}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 \mathbf{u}_d}{\partial x \partial z} \right)^2 + \left(\frac{\partial^2 \mathbf{u}_d}{\partial y \partial z} \right)^2 \right] \right\} dx dy dz \tag{2.4}$$

where x, y, z refer to locations within the domain Ω_y, bounded by (0,0,0) and (X, Y, Z).

More advanced prior distributions would be representative of a population of interest, such as Ashburner et al. demonstrate for affine registration [15]. Such a prior could be estimated for non-rigid registration from many registrations, but it may be computationally expensive to use due to the extensive covariance structure, although it would clearly be beneficial. The principal modes of variation, as obtained by principal component analysis using the covariance matrix of many registrations, have been used as a basis set for computationally efficient registration [104]. A symmetric prior was proposed by Ashburner et al. [8] and extended to 3-D [9] for a high dimensional finite-element approach to registration. Their prior provides equivalent penalties for the forward, and backwards transformation, which is not the case with most priors. Adaptive spatial priors could be inferred from the registration of image pairs, or groupwise registrations. This could draw from work in other aspects of brain imaging [62][78][72], and is further discussed in chapter 9.

Alternative Regularisation Methods

Some fluid type registration approaches such as demons [180] and diffeomorphic demons [186], provide an ad-hoc approach to regularising the mapping between images. These approaches regularise through the use of a Gaussian kernel convolution with the deformation, or velocity fields. This constrains all the deformations to be of a certain scale. Furthermore, mappings may be constructed by composing multiple transformations consecutively, where each of the individual transformations may be separately smoothed using an additional Gaussian kernel. The selection of an appropriate Gaussian is unintuitive, and changes to the regularisation scheme have a very strong influence on the registration.

An additional problem with these approaches is that they do not optimise the complexity of the overall mapping at the same time as encouraging spatial smoothness, as in LDDMM [22]. Instead, these aspects are decoupled into separate terms, which makes it computationally efficient, but does not directly optimise the intended gradient. This may result in a poor constraint that infers inaccurate mappings.

The next section describes some image similarity measures that are used in non-rigid registration.

2.1.4 Image Similarity Measures

A generative model for registration can be formulated such that an estimate of the observed target image is generated from the transformed source image, as given in equation 2.2. Additionally, this estimate will contain errors due to mismatched structures, or image noise. As such, a random noise term is contained in the model. The most common approach is to assume additive independent, and identically distributed Gaussian noise. This leads to a model of the form:

$$\mathbf{y} = \mathbf{t}(\mathbf{x}, \mathbf{w}) + \mathbf{e} \tag{2.5}$$

where $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$, $\mathbf{I}$ is an identity matrix, and σ^2 is the noise variance. The likelihood, which is a probabilistic measure of model fit, for this model is related to the commonly used similarity measure, sum of squared differences (SSD):

$$C_{sim} = (\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w}))^2 \tag{2.6}$$

The limitation of SSD is that both images must have identical contrast to be reasonably compared. This can be rectified by explicitly modelling an intensity mapping of one image to match that of the other [63][10][5]. Such an intensity mapping E, parametrised by $\mathbf{b}$, $E(\mathbf{b}; \mathbf{y})$, can be substituted into the generative model for $\mathbf{y}$. E could be used to provide a linear intensity mapping:

$$E(\mathbf{b}; \mathbf{y}) = b\mathbf{y} \tag{2.7}$$

where b is a single variable that scales the image. Non-linear intensity mappings give a greater degree of flexibility, and can be defined as:

$$E(\mathbf{b}; \mathbf{y}) = \sum_{j=1}^{n} b_j \mathbf{y}^{j-1} \tag{2.8}$$

where an nth order mapping is used. The use of an intensity mapping corrects for differences in contrasts between the two images due to the image acquisition. If significant bias fields are present, these mapping can be permitted to vary smoothly over space [13] to jointly model the difference in image contrast and bias fields [5]. The intensity mapping parameters and transformation parameters together describe the generative model, and can be inferred simultaneously from

the data when using the Bayesian inference techniques that will be described later in this thesis. Incorporating such a mapping makes the use of a generative model flexible enough for non-rigid registration of real MR data.

Alternative generative models have been proposed that permit different image similarity measures. For example, Zöllei et al. provide an interesting registration formulation through modelling the joint image intensity statistics as a multinomial distribution [208]. This allows for a multi-modal registration cost-function, which minimises the entropy of the joint image intensity distribution. Additionally, this provides a mechanism for informative priors on the similarity function, which correspond to known intensity-mappings between image modalities.

Alternative Image Similarity Measures

Registration approaches that are not based on an underlying generative model allow an arbitrary choice of similarity function. Popular alternative similarity function choices include measures derived from information theory, such as mutual information (MI) [187][119][156]. MI makes no assumptions regarding the relationship between image intensities, but such flexibility is not necessary for single-modal MR images. MI is commonly calculated from the joint histogram of image intensities. This requires the width of each intensity interval, or bin, to be selected for each image. This needs to be considered to avoid intensities related to different tissue types falling into the same bin, and being counted as equivalent. A further common measure is cross-correlation, which measures the linear dependency between the intensities in two images [19][68]. Such an approach is equivalent to estimating a linear intensity mapping between images. The correlation ratio allows an arbitrary functional relationship between the intensities of the two images [149]. Probabilistic interpretations of all of these cost functions was investigated in Roche et al. [148].

Mutual information has been formulated as conditional on the position in image space to compensate for bias fields [116]. Cross-correlation has been extended similarly [83][17] by calculating statistics only on a local region of an image. These approaches help compensate for bias fields as the local intensity heterogeneities are accounted for. However, there is a potential for the similarity metric to overfit the data, e.g. give a high similarity for different tissue types, as the local statistics do not contain sufficient information. This is because these approaches do not model a smoothly varying intensity difference, but instead calculate the similarity measure based on the statistics of a local region of arbitrary size.

In terms of single modal registration of real MR images of the brain, such as considered in this thesis, SSD, with an appropriately modelled intensity relationship, should provide at least as effective a cost-function, if not better than any other measure. This similarity measure also fits within the generative model framework.

2.1.5 Hierarchical Registration Schemes

Hierarchical registration schemes provide a framework for efficient and accurate registration between images [115]. There are two hierarchical registration strategies that are typically considered: those that manipulate the information content of images, referred to as *multi-resolution* frameworks [18], and those that consider the complexity of the transformation, referred to as *multi-level* registration frameworks [160]. Note that for transformation models that do not use a sparse parametrisation, multi-resolution schemes are intrinsically multi-level.

Multi-resolution, at its simplest, is the registration of subsampled images to one another, in a coarse-to-fine fashion. This is commonly accompanied by image smoothing, the level of which may be based on a scale-space interpretation [191] that ensures image features of a particular scale are visible in the data. Smoothing the data by a greater amount at coarser scales allows the larger scale image features to be matched first. The reduction of small scale image features reduces the choices of voxel matching locations dramatically, and helps ensure the registration method can find a reasonable solution. Subsampling also improves the computational efficiency, as less voxels are present at the coarser levels.

Multi-level schemes follow in a similar manner. Initially a relatively coarse transformation is used to estimate the large scale image deformations, followed by successively increasing the number of degrees of freedom at finer levels. As with multi-resolution schemes, this leads to a large improvement in computational efficiency, and reduces the risk of finding suboptimal solutions. In most multi-level approaches, a single set of transformation parameters, $\mathbf{w}$, is estimated. These initially describe a coarse transformation, but are refined at each multi-level. As such, $\mathbf{w}$ will describe the entire deformation.

Multi-resolution and multi-level schemes are often used together, and for many approaches to registration they are equivalent. In this thesis they will jointly be considered as a hierarchical registration scheme.

2.1.6 Optimisation of Registration

Once an appropriate similarity term and regularisation criterion have been established, the process of finding the optimal set of parameters to fit the model needs to be considered. A wide variety of optimisation procedures have been used in non-rigid registration; however, for brevity only a brief description of the two most common strategies is presented.

Gradient Descent

Gradient descent is a first-order optimisation algorithm, and is one of the simplest optimisation schemes. The gradient of the registration cost function is evaluated with respect to each transformation parameter independently [156][22]. The gradient vector is evaluated for the cost function from the current set of parameters by:

$$\nabla C(\mathbf{w}, \phi, \lambda) = \frac{\partial C(\mathbf{w}, \phi, \lambda)}{\partial \mathbf{w}} \tag{2.9}$$

where C is the cost function given by equation 2.3. Once the gradient has been calculated, a line search is performed along $\nabla C(\mathbf{w}, \phi, \lambda)$ to choose the next set of parameters. Gradient descent has been characterised as having slow convergence in non-rigid rigid registration problems when compared to using a Gauss-Newton type scheme, and is more susceptible to locally optimal solutions [206].

Gauss-Newton

Gauss-Newton is a second order optimisation algorithm that is used to solve problems of the non-linear least squares form. This problem is non-linear in that the output of the model is given by a non-linear combination of the model parameters. As with gradient descent, the gradient of the cost function with respect to the model parameters is required and is calculated by:

$$\nabla C(\mathbf{w}, \phi, \lambda)_j = \frac{\partial C(\mathbf{w}, \phi, \lambda)}{\partial \mathbf{w}_j} = \sum_{i=1}^{N_v} \phi(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w}))_i \frac{\partial \mathbf{t}(\mathbf{x}, \mathbf{w})_i}{\partial \mathbf{w}_j} + \lambda \frac{\partial C_{reg}(\mathbf{w})}{\partial \mathbf{w}_j} \tag{2.10}$$

where j indexes any transformation parameter, and i indexes through the N_v voxels. This can be re-written using linear algebra.

$$\nabla C(\mathbf{w}, \phi, \lambda) = \phi \mathbf{J}^T \mathbf{k} + \lambda \Lambda \mathbf{w} \tag{2.11}$$

where $\mathbf{J}$ is the first-order matrix of partial derivatives of the model parameters with respect to the transformed source image, $\mathbf{k}$ is the residual error of the model defined as $\mathbf{y} - \mathbf{t}(\mathbf{x} - \mathbf{w})$. $\Lambda\mathbf{w}$ provides the gradient of the regularisation, as specified by the inverse covariance regularisation matrix, Λ.

Additionally, the matrix of second-order partial derivatives of the generative model is required. This is known as the Hessian matrix. In Gauss-Newton, the Hessian is approximated by assuming that the second order derivatives are small enough to be negligible. This is then equivalent to a 1st order Taylor series approximation of the transformation model. This approximation is commonly used in image registration, both for computational reasons and because the second derivative can reduce the stability of the optimisation [138]. The approximate Hessian entry for a given pair of transformation parameters $\mathbf{w}_k$ and $\mathbf{w}_l$ is given by:

$$\mathbf{H}_{j,l} = \frac{\partial^2 C(\mathbf{w}, \phi, \lambda)^2}{\partial \mathbf{w}_j \partial \mathbf{w}_l} = \sum_{i=1}^{N_v} \phi \frac{\partial \mathbf{t}(\mathbf{x}, \mathbf{w})_i}{\partial \mathbf{w}_l} \frac{\partial \mathbf{t}(\mathbf{x}, \mathbf{w})_i}{\partial \mathbf{w}_l} + \lambda \frac{\partial C_{reg}(\mathbf{w})}{\partial \mathbf{w}_j} \frac{\partial C_{reg}(\mathbf{w})}{\partial \mathbf{w}_l} \quad (2.12)$$

and again, can be re-written using linear algebra as:

$$\mathbf{H} = \phi \mathbf{J}^T \mathbf{J} + \lambda \Lambda \quad (2.13)$$

Each iteration consists of calculating the gradient and the Hessian given the current set of parameters. The next set of parameters is calculated by:

$$\mathbf{w}_{new} = \mathbf{w}_{old} - \mathbf{H}^{-1} \nabla C(\mathbf{w}, \phi, \lambda) \quad (2.14)$$

The advantages of using Gauss-Newton over gradient descent is that the Hessian provides a measure of the distance a cost function derivative can be trusted for each parameter, rather than performing a search along the line of the gradient. Additionally, the Hessian recognises those parameters that covary, and thus should move together. This allows the gradient to be rotated appropriately. Thus Gauss-Newton requires fewer iterations than gradient descent, and may be more resistant to local minima. The inverse of the Hessian can also be interpreted as the covariance matrix of $\mathbf{w}$, which will be revisited in the next chapter.

2.1.7 Summary

As this section has illustrated, a generative model can be formulated to provide an accurate and flexible approach to medical image registration between MR images of the human brain. The advantage of using a generative approach is that it allows a probabilistic model description. Model parameters can be regularised through the use of prior distributions, which probabilistically describe the set of possible values they would be expected to take. The optimisation, or inference, of these model parameters should incorporate the prior distributions in a principled fashion. Furthermore, the inference should take advantage of the probabilistic nature of the model, and allow distributional estimates of the inferred model parameters. Bayesian statistical inference provides the only coherent framework for the adjustment of belief (in the form of probability density functions) in the presence of new information [41]. The theory and use of Bayesian inference is explored in the next section.

2.2 Bayesian Inference

Bayesian statistics provides a principled mechanism for the incorporation of prior information regarding the nature of unknown model parameters, when inferring the optimal model parameters. A further advantage of using a probabilistic inference scheme, is that it also facilitates, and intrinsically considers, estimates of parameter uncertainty. Bayesian inference schemes have been previously applied in medical image registration with methods including [66][67][10][13][2][183][146][14].

2.2.1 Bayes' Rule

Bayes' rule provides a mechanism for the inference of a set of parameters Θ, of a given model $\mathcal{M}$, given some observable data $\mathcal{D}$. Bayes' rule allows the incorporation of prior beliefs about the parameter values, $P(\Theta)$, and infers a posterior probability distribution on the model parameters given the data and the prior. Bayes' rule states:

$$P(\Theta|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\Theta,\mathcal{M})P(\Theta|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})} \tag{2.15}$$

The *posterior* probability of the model parameters, $P(\Theta|\mathcal{D},\mathcal{M})$ is given in terms of the *likelihood* of the data given the model parameters, $P(\mathcal{D}|\Theta,\mathcal{M})$, multiplied

by the *prior* probability of the parameters, $P(\Theta|\mathcal{M})$, and normalised by the model *evidence*, $P(\mathcal{D}|\mathcal{M})$. The model evidence is not dependent on Θ, and can therefore be calculated by integrating the likelihood over the set of model parameters. This gives rise to the following relationship:

$$P(\Theta|\mathcal{D}, \mathcal{M}) \propto P(\mathcal{D}|\Theta, \mathcal{M})P(\Theta|\mathcal{M}) \tag{2.16}$$

In subsequent equations, the model $\mathcal{M}$ will be implicit for clarity.

2.2.2 Priors

The use of a prior distribution on the set of model parameters, $P(\Theta)$, is part of Bayesian inference. Priors range from being entirely uninformative, e.g. a uniform distribution across all possible values, to highly specific delta functions. Priors need to embed any *a priori* knowledge of the model parameters.

Hierarchical priors can be constructed that allow for a flexible approach to the specification of the prior distribution. In such a hierarchical model, the parameters of the prior are modelled as random variables that can be inferred from the data. This allows a more "objective" approach to inference, as the parameters are estimated from the data [69], rather than subjectively selected.

Such a hierarchical model requires hyperparameters to be defined as random variables. These are variables that do not directly contribute to the likelihood, but instead affect the priors on other parameters. Regularisation strength is an example of a hyperparameter. Hyperparameters additionally require a prior distribution to be specified on their value, and themselves follow a distribution.

Hierarchical priors have been demonstrated to provide regularisation in brain imaging, for example in fMRI detection [196][135] and segmentation [195].

2.2.3 Inference

Full Bayesian inference consists of finding the posterior distribution of parameter values according to Bayes' rule, as described in equation 2.15. However, a difficulty is presented in this equation: the calculation of the model evidence, $P(\mathcal{D})$, which is required to normalise the posterior distribution $P(\Theta|\mathcal{D})$, is often intractable to calculate as it requires integrating the likelihood over the set of all model parameters. Additionally, to calculate the marginal posterior distribution

of an individual model parameter requires integration over the joint posterior distribution of all other model parameters.

Markov Chain Monte Carlo

One option is to estimate the required integrals numerically using tools such as Markov Chain Monte Carlo (MCMC) [24]. MCMC methods attempt to numerically build-up an estimate of the true posterior parameter distribution by sampling sets of parameter values. These samples are drawn in a random walk through parameter space, where the current sample is correlated only with the preceding sample. Samples of a point in parameter space are kept according to the unnormalised posterior probability, given in equation 2.16. This means that given sufficient samples, the true posterior distribution can be constructed, and no approximations need to be made regarding the distribution of the parameters. Several MCMC algorithms exist, the most popular include: Metropolis-Hastings [122][81] and Gibbs sampling [70]. The most appropriate algorithm depends on the problem of interest. MCMC methods are particularly computationally demanding for problems which have a large number of model parameters, such as high resolution image registration. This is because of the need to sample the full joint distribution of parameters. However, MCMC has been demonstrated as applicable in 2-D, or low resolution non-rigid registration [66][146].

Approximate Bayesian Inference

As numerical integration is particularly costly, the majority of Bayesian approaches to image registration tend to not perform full Bayesian inference. Instead, point estimates of model parameters can be inferred using the relationship given in equation 2.16. As point estimates are inferred, integration is not required, thereby making such methods computationally efficient. This approach is referred to as *maximum-a-posteriori* (MAP), as it aims to calculate the most probable value of the posterior distribution of the model parameters. Many approaches to image registration have used MAP, and include: Gee et al. [65], Andersson et al. [5] and the various methods of Ashburner and Friston [10][13][7][14].

MAP does not allow for inference on hierarchical probabilistic models, and only provides point estimates of registration parameters. On the other hand, numerical integration using MCMC is computationally too expensive for image registration. Instead, an alternative option is to consider simple analytic approximations to the posterior distribution. These allow for full Bayesian inference while

being computationally more efficient than numerical approaches. Allassonniére et al. [2] use such a scheme to infer parametric distributions for variables using a variant of the expectation-maximisation (EM) algorithm [48]. Van Leemput [183] uses the Laplace approximation and an EM scheme for inference of model parameters.

In this thesis, mean-field variational Bayesian (VB) [99][16][95] is explored as an approximate solution for full Bayesian inference, although other approaches such as restricted maximum likelihood (REML) [133], could have been chosen.

2.2.4 Variational Bayes

The mean-field VB approach for inference of graphical models [99][16][95] builds on previous work using the mean-field approximation [137]. VB is related to the popular expectation-maximisation (EM) approach [48] and is sometimes referred to as variational EM in the literature. VB allows tractable Bayesian inference of an approximate posterior probability distribution of the model parameters. This is achieved by approximating the posterior parameter distribution $P(\Theta|\mathcal{D})$ as a simpler, parametric probability distribution function, $q(\Theta)$:

$$q(\Theta) \approx P(\Theta|\mathcal{D}) \tag{2.17}$$

Variational Free Energy

The negative variational free energy, $\mathcal{F}$, is the cost function of variational Bayesian inference. $\mathcal{F}$ can be derived using the log model evidence, which is written as:

$$\log P(\mathcal{D}) = \int \log \frac{P(\mathcal{D},\Theta)}{P(\Theta|\mathcal{D})} d\Theta \tag{2.18}$$

As VB approximates the posterior parameter distribution as $q(\Theta)$, the log evidence can be re-written by taking the expectation with respect to $q(\Theta)$:

$$\begin{aligned} \log P(\mathcal{D}) &= \int q(\Theta) \log \frac{P(\mathcal{D},\Theta)}{P(\Theta|\mathcal{D})} d\Theta \\ &= \int q(\Theta) \log \frac{P(\mathcal{D},\Theta) q(\Theta)}{P(\Theta|\mathcal{D}) q(\Theta)} d\Theta \\ &= \int q(\Theta) \log \frac{P(\mathcal{D},\Theta)}{q(\Theta)} d\Theta + \int q(\Theta) \log \frac{q(\Theta)}{P(\Theta|\mathcal{D})} d\Theta \\ &= \mathcal{F} + \mathcal{KL} \end{aligned} \tag{2.19}$$

where $\mathcal{F}$ is the negative variational free energy and $\mathcal{KL}$ is the Kullback-Leibler distance [108] between the approximate, and the theoretical posterior distribution of Θ.

$$\mathcal{KL} = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|\mathcal{D})} d\Theta \tag{2.20}$$

$\mathcal{KL}$ seems like an ideal quantity to minimise, as it directly measures the distance of the approximate solution from the true theoretical distribution. However, this necessitates already knowing $P(\Theta|\mathcal{D})$, which is not the case. Fortunately, $\mathcal{KL}$ is always positive, which makes $\mathcal{F}$ a lower bound on the log-evidence. Closer examination of $\mathcal{F}$ shows that it can broken up into two composite components.

$$\begin{aligned} \mathcal{F} &= \int q(\Theta) \log \frac{P(\mathcal{D}, \Theta)}{q(\Theta)} d\Theta \\ &= \int q(\Theta) \log \frac{P(\mathcal{D}|\Theta)P(\Theta)}{q(\Theta)} d\Theta \\ &= \int q(\Theta)(\log P(\mathcal{D}|\Theta) d\Theta + \int q(\Theta)(\log P(\Theta) - \log q(\Theta)) d\Theta \\ &= \mathcal{L}_{av} - D_{KL}(q(\Theta) \| P(\Theta)) \end{aligned} \tag{2.21}$$

$\mathcal{L}_{av}$ is the marginal value of the log likelihood with respect to the approximate posterior distribution, $q(\Theta)$:

$$\begin{aligned} \mathcal{L}_{av} &= \int q(\Theta)(\log P(\mathcal{D}|\Theta)) d\Theta \\ &= < \log P(\mathcal{D}|\Theta) >_{q(\Theta)} \end{aligned} \tag{2.22}$$

where the angled brackets correspond to an expectation with respect to the subsequent subscripted term. The $D_{KL}(q(\Theta)\|P(\Theta))$ term can be re-written like so:

$$\begin{aligned} -D_{KL}(q(\Theta)\|p(\Theta)) &= \int q(\Theta)(\log p(\Theta) - \log q(\Theta)) d\Theta \\ &= -\int q(\Theta) \log \frac{q(\Theta)}{p(\Theta)} d\Theta \end{aligned} \tag{2.23}$$

which is recognisable as the Kullback-Liebler distance between the approximate posterior, and the prior distributions.

Intuitively, $\mathcal{F}$ provides a summary of model fit given the approximate posterior distribution, whilst penalising the deviation of $q(\Theta)$ from the prior, $P(\Theta)$.

Mean-Field Approximation

The mean-field approximation is the assumption of independence between sets of parameters in the approximation to the posterior distribution [137]:

$$q(\Theta) = \prod q_{\Theta_i}(\Theta_i) \tag{2.24}$$

where the parameters are collected into separate groups Θ_i, each of which have an approximate marginal posterior distribution $q(\Theta_i)$. In general, a full factorisation of parameters is not required [16]. Instead, independence can be assumed between sensible parameter groupings; e.g. transformation parameters, and noise parameters.

The assumption of parameter group independence means that during inference each parameter group is inferred conditionally on the approximate posterior distributions of the other parameters. As such, an iterative inference procedure is required to account for any real relationships between the parameter groups that are described as independent according to the mean-field approximation. This means that the inferred posterior distributions can still be related, but this relationship is not modelled explicitly.

The form of the factorised posterior distributions, $q(\Theta_i)$, is not fixed arbitrarily, but is instead determined algebraically by combining the likelihood expression with the prior distribution.

Calculus of Variations

Variational calculus can be used to derive analytic iterative updates for each approximate posterior distribution with the aim of maximising $\mathcal{F}$.

The negative variational free energy, $\mathcal{F}$ can be described as a functional, f, which is a function of a function.

$$\mathcal{F} = \int f(\Theta, q(\Theta), q'(\Theta)) d\Theta \tag{2.25}$$

where $q'(\Theta)$ corresponds to the first differential of $q(\Theta)$ with respect to Θ.

The calculus of variations provides a framework to maximise functionals. As the parameter groups are independent, $\mathcal{F}$ only needs to be maximised with respect to a certain parameter group at a time, Θ_i, where all the other parameters $\Theta_{\not i}$ are constant.

This allows a new functional, g, that only relates to a particular parameter group, i: $\int g(\Theta_i, q(\Theta_i), q'(\Theta_i))d\Theta_i = \mathcal{F}$, where:

$$\begin{aligned} g(\Theta_i, q(\Theta_i), q'(\Theta_i)) &= \int f(\Theta, q(\Theta), q'(\Theta))d\Theta_{\not i} \\ &= \int q(\Theta_{\not i})(\log P(\mathcal{D}|\Theta) + \log P(\Theta) - \log q(\Theta))d\Theta_{\not i} \end{aligned} \tag{2.26}$$

The Euler-Lagrange differential equation can be formulated to find the point where the functional derivative, g, is zero for a parameter set Θ_i. This can be written as:

$$\frac{\partial g(\Theta_i, q(\Theta_i), q'(\Theta_i))}{\partial q(\Theta_i)} - \frac{d}{d\Theta_i}\frac{\partial g(\Theta_i, q(\Theta_i), q'(\Theta_i))}{\partial q'(\Theta_i)} = 0 \tag{2.27}$$

In this case, the second term of the equation is equal to zero, as g, given in equation 2.26, does not depend on $q'(\Theta)$. Subsituting equation 2.26 into 2.27 gives:

$$0 = \frac{\partial}{\partial q(\Theta_i)} \int q(\Theta_{\not i})(\log P(\mathcal{D}|\Theta) + \log P(\Theta) - \log q(\Theta))d\Theta_{\not i} \tag{2.28}$$

This expression provides the point where the derivative of g, hence $\mathcal{F}$ with respect to Θ_i, is zero. Therefore, the approximate posterior distribution, $q(\Theta_i)$, that satisfies this local optima can be found by rearrangement and writing the marginalisation as an expectation:

$$\log q(\Theta_i) =< \log P(\mathcal{D}|\Theta) + \log P(\Theta) >_{q(\Theta_{\not i})} + \kappa \tag{2.29}$$

where κ is a constant. The integral on the left hand side disappears because $q(\Theta_i)$ is independent of $q(\Theta_{\not i})$. $\log q(\Theta_i)$ can now be chosen to match the right hand side of the equation. For certain likelihood models and approximate posterior distributions, the parameters of $q(\Theta_i)$ can be found analytically by algebraic manipulation.

It should be noted that although the variational free energy is the measure being optimised in this procedure, it does not ever need to be calculated explicitly. However, it could in principle be used to measure convergence, or perform model comparison.

Non-Linear Models

VB is is not tractable for arbitrary non-linear forward models. To account for non-linear models, such as those required in non-rigid registration, a linear approximation to the transformation model must be used. The forward model can be approximated as linear using a first [34], or second [195] order Taylor series approximation.

2.2.5 Summary

The use of VB provides a principled, and tractable framework for inference on hierarchical probabilistic models; therefore, VB can be used to infer an appropriate balance of regularisation priors, and data fidelity based on the data. Furthermore, VB facilitates the inference of approximate posterior parameter distributions, providing a measure of parameter uncertainty. The final section of this chapter provides details on how non-rigid registration can be used in the derivation of morphological biomarkers that are associated with AD.

2.3 Morphometric Biomarkers

As was described in section 1.2.3, structural MRI of the brain is well suited to the analysis of morphological variability of neuroanatomical structures. The identification and measurement of consistent patterns of atrophy and morphological changes that are indicative of progression to, or of, AD have a variety of applications, as previously described in section 1.2.1. This section provides an overview of some common approaches used in the extraction of morphological biomarkers of AD. There are two main groups of automated methodological approaches for identifying morphological biomarkers: voxel and deformation-based morphometry [11].

2.3.1 Voxel Based Morphometry

Voxel-Based Morphometry (VBM) provides a tool for analysing voxel-wise differences in tissue properties across a range of subjects [199]. VBM requires a probabilistic segmentation map of a certain tissue type, e.g. grey matter, which indicates the probability that a given voxel contains a tissue type. Probabilistic

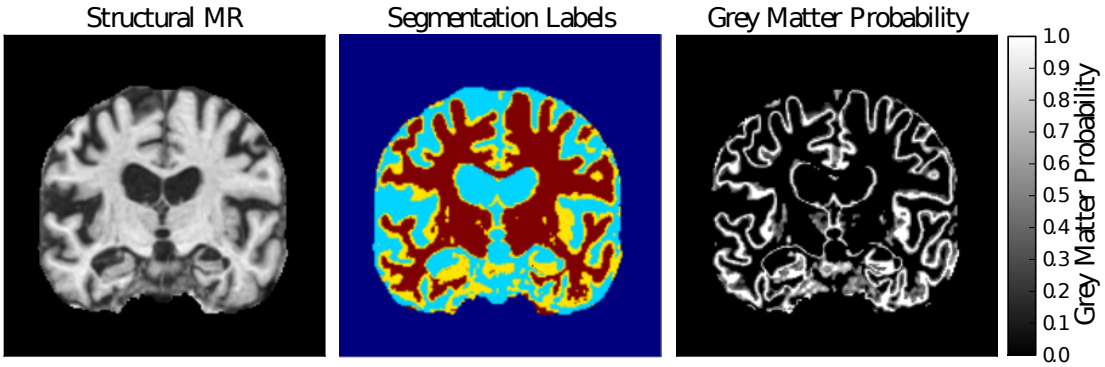

Figure 2.2: An example of the probabilistic segmentation maps required for VBM as estimated by FAST. The original image is given on the left, the middle image indicates the most likely tissue class for different regions, where different colours correspond to different tissue types. The right image shows the grey matter probability map.

segmentation maps can be automatically defined through a model-based segmentation, and an example map is given in Figure 2.2 as segmented by FAST [204]. These maps are then registered to an atlas space. The transformed tissue probability map is modulated (multiplied) by the determinant of the warp field Jacobian, to compensate for the expansion/contraction of voxels [71]. In traditional VBM approaches, statistical tests are performed in a voxel-wise manner to find regions of significant differences between subjects. However, these registered and modulated segmentation maps could be examined using a variety of statistical tools with the aim of locating consistent biomarkers of disease. Grey matter (GM) probability maps have previously been used in the discrimination of AD [106], while other methods have incorporated white matter (WM) and cerebrospinal fluid (CSF) as well [120][184], see [44] for a comparison of classification methods using VBM features.

Role of Registration

Spatial normalisation in VBM is essential in the generation of useful feature data. Commonly, it is used to only resolve the large scale differences in brain morphology, leaving spatially localised differences intact. This level of registration accuracy preserves differences between the subject groups. If the registration produced a perfect correspondence in terms of tissue properties, then there would be no differences in tissue maps between subjects to evaluate, and all the discriminative information would be encoded in the deformation field. The use of the deformation Jacobian determinant to modulate the tissue map leads to a

continuum between voxel and tensor based morphometry (described in the next section), allowing either coarse or accurate registration to yield informative features. An appropriate level of accuracy for VBM can be achieved with a relatively coarse transformation, e.g. a 10mm control point spacing using a B-spline FFD model.

Modern approaches to VBM spatially normalise the segmented tissue probability maps to a tissue probability atlas, which leads to any results being directly attributable to differences in that tissue class. Previous approaches used registration of the MR image directly, which elicited some controversy [27] due to the sensitivity of VBM under imperfect registration, giving results that are not necessarily caused by differences in a particular tissue class [12]

2.3.2 Deformation- and Tensor-Based Morphometry

Deformation-, and tensor-based morphometry allow for the assessment of differences in brain morphometry between two images using features derived from the deformation field obtained from non-rigid registration of the images. Deformation-based morphometry is used to "identify differences in the relative positions of structures within the subjects' brains", whereas tensor-based morphometry refers to the measurement of local shape differences in brain structures [11]. This thesis focuses on tensor based morphometry (TBM), as it has been more commonly used in the investigation of neurodegenerative disease because of its ease of interpretation.

Tensor-based morphometry takes its name from the analysis of the Jacobian tensor of the deformation field, which is estimated from the non-rigid registration of two images. TBM provides a framework to summarise informative features from a deformation field. The most common feature to be used is the determinant of the Jacobian tensor, or its log, which provides a measure of the expansion or contraction of a voxel due to the inferred deformation [38]. Ridgway [141] provides an overview of various other TBM features for use in the detection of AD, and finds that the log determinant of the Jacobian tensor yields the largest distinction between population groups. Taking the log of the Jacobian determinant is likely to prove more useful in statistical analysis, as the data distribution will be symmetric and closer to normal [111]. This symmetry is useful in statistical analysis, as a doubling of volume or a contraction of 50% will be equally spaced from a level of constant volume. This allows standard distance measures to be used without bias towards contraction or expansion.

There are two slightly different manners in which TBM can be applied: for the evaluation of differences between a subject and an atlas, or to describe the changes in a subject across time.

Atlas Based Tensor-Based Morphometry

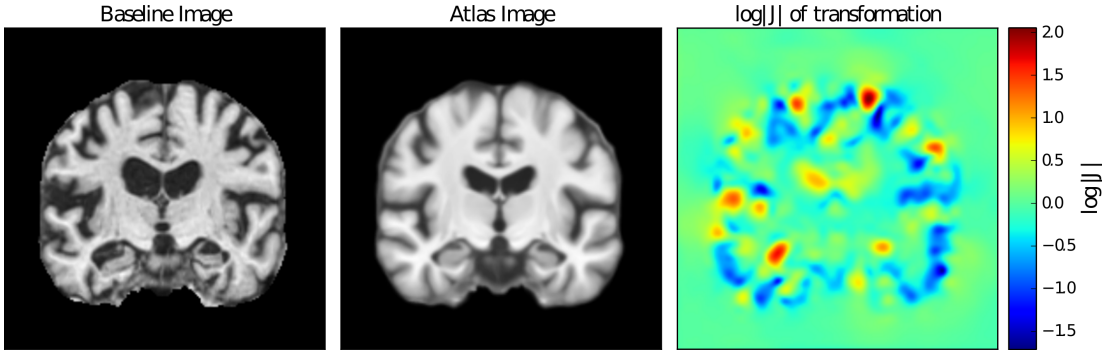

Figure 2.3: An example of atlas based tensor based morphometry features derived from the registration of a subject image (left) to the atlas image (middle). The TBM features (right) shows the log determinant of the transformation, and indicate the level of expansion or contraction at each spatial location. This subject suffers from AD, and so the rapid expansion of CSF spaces and loss of grey matter is visible.

TBM, using the determinant, or log determinant of the Jacobian tensor, has been frequently used to measure differences in the size of brain structures between atlas and subject images. By analysing the TBM features between and across groups, regions of consistently differing sizes can be evaluated for use as biomarkers of pathology. This is similar in principle to the use of volumetry, which is the measurement of the volume of specific brain structures, except no segmentations or regions of interest are required, as the deformation field encodes the relative size difference of each voxel from an atlas image [38]. This presents an advantage over volumetry as changes can be measured on a voxel level, providing the registration is sufficiently accurate. Atlas based TBM has been explored in the analysis of AD [179][113][90]. Conversely, it has also been used in the analysis of brain growth [1]. An example of atlas based TBM is given in Figure 2.3.

Longitudinal Tensor Based Morphometry

Longitudinal imaging of subjects with Alzheimer's disease has been used to provide estimates of the rate of brain atrophy that are a hallmark of progressive

dementia. Previous approaches that provide a global measure of the rate of atrophy have been developed, which only require affine registration: the boundary shift integral [59] and SIENA[172]. However, these methods have a limited ability to provide a spatially localised estimate of the rate of atrophy which is necessary for differentiating between pathologies.

Non-rigid registration of longitudinal MRI scans of the brain has been used to estimate the level of atrophy between two images [60], and such methods have been shown to be sensitive to anatomical changes in asymptomatic subjects with AD [159]. An example of longitudinal TBM is given in Figure 2.4.

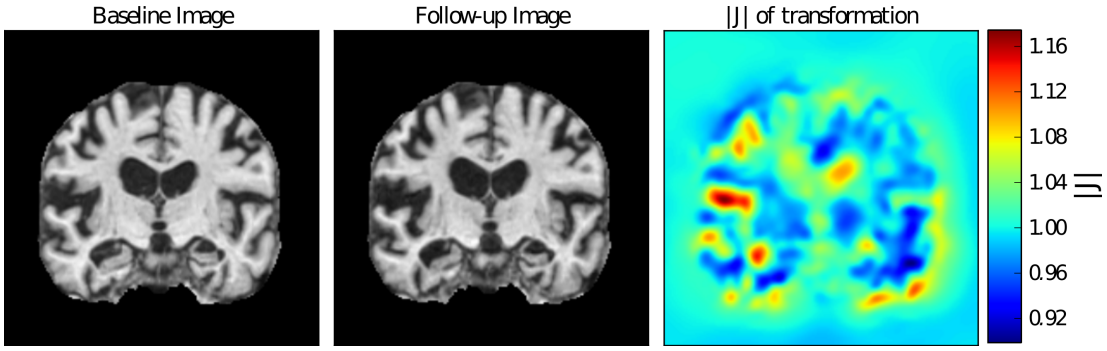

Figure 2.4: An example of longitudinal tensor based morphometry features derived from the registration of the follow-up image (middle) to the baseline image (left). The TBM features (right) show the determinant of the transformation, and indicate the level of expansion or contraction at each spatial location. This subject suffers from AD, and so the rapid expansion of CSF spaces and loss of grey matter is visible.

Role of Registration

As TBM features are entirely derived from non-rigid registration, they are sensitive to differences in the registration methods used to estimate them. This is illustrated in Figure 2.1, where different levels of regularisation are used for registering a subject to an atlas. Clearly, the inference of biologically plausible Jacobian maps will be dependent on the level of regularisation.

It is difficult to evaluate which registration model, and parametrisation, generates the most realistic TBM features, as ground truth deformation fields are unknown. Camara et al. [32] proposed a simulation based approach that provides gold standard deformation fields, which can be used to measure the accuracy of the tensor features. They found that a B-spline FFD transformation model produced a more accurate estimate of atrophy than fluid based registration. However,

this evaluation is limited by the biological accuracy of the simulations. Yanovsky et al. [200] provide a systematic comparison of several in-house registration methods for creating Jacobian maps. This evaluation is based on serial registration of ten two week follow up scans, where no changes are expected, and ten one year follow-up scans, where changes are expected. As there are no ground truth deformation fields, or Jacobian features, only the consistency and magnitude of changes between registering baseline to atlas, and vice versa, can be evaluated. Furthermore, the optimal level of regularisation may vary across subjects, because of the large morphological differences in different images. This problem could be approached by inferring the level of regularisation in image registration, which will be described in chapter 3 and demonstrated in chapter 7.

2.3.3 Atlas Construction

Standard atlas spaces exist that can be used as a common reference frame. However, for many tasks an atlas that is more representative of the population of interest is preferable. The use of a sharp and and representative population atlas is a key issue for spatially normalised statistical analysis. Single subjects can be selected for use as a standard co-ordinate system. This has limited application as the complex inter-subject anatomical variability cannot be encapsulated from a single image, and therefore in any analysis there will be an inherent bias towards the atlas subject. Several frameworks have been proposed to reduce the bias in atlas estimation. These range from iterative averaging approaches [75], to groupwise registration approaches using small [23][174][209], or large [100] deformation transformation models. More complex approaches to modelling the populations of an atlas have described. Allassonniére et al. estimate a continuous multi-modal atlas [2], Blezek and Miller use the mean-shift algorithm to find the modes of the image population [25] and Sabuncu et al. use a Gaussian mixture model to find the modes of an image population [157]. However, it is unclear how multiple atlases can be best used in practise.

The iterative atlas creation framework proposed by Guimond et al. [75] is preferred in this thesis: this is because of its capacity to work with a pairwise registration algorithm, rather than necessitating the use of a groupwise registration approach. For a given group of subject images the framework proceeds as follows:

1. The images are affinely aligned to any atlas space (e.g. MNI 152).

2. An initial atlas estimate is calculated by averaging the subject images.

3. Each subject image is non-rigidly registered to the current atlas estimate.

4. The registered subject images are averaged to make a mean intensity image.

5. The deformation fields used to transform the subject to the atlas are inverted, and then averaged to find the mean inverse deformation.

6. The mean inverse deformation is applied to the mean intensity image to generate a new atlas estimate.

7. Repeat from step 3 until subsequent atlas iterations converge (differ by less than 1%).

In practise, this algorithm converges in four, or five iterations, and is minimally biased towards the MNI 152 atlas, to which the images are initially affinely registered.

This procedure fits the observed data, the population of images, to the estimated population average, which whould ideally be a part of the model. A more principled framework, would iteratively estimate the population average by registering the current atlas estimate to the population of images. This would form a more reasonable generative model of the observed images, and the image gradients would be smoother. Unfortunately, this requires inverting the deformation field to resample the subject images in the average space. Inverting transformations may be slow and unreliable for certain transformation models. Future work will investigate how such a procedure could be integrated with the work in this thesis.

2.3.4 Statistical Prediction

Once an appropriate set of biomarkers have been extracted from the data and normalised to an appropriate reference frame, they can be statistically analysed across subject groups. This analysis takes the form of a multivariate data classification, or regression problem. Through these methods, an outcome variable, e.g. disease group or score, can be estimated from a set of feature data [24], e.g. VBM or TBM maps. This thesis examines the classification of subjects into disease groups based on image derived measures of brain morphometry. Such an approach allows the objective diagnosis, or, if the outcome variable corresponds to future disease status, prognosis of a subject. Classification allows the predictive

properties of the data to be quantified in terms of how accurately it can be used to differentiate between disease groups. This allows the discriminative quality of different feature types, e.g. VBM or TBM, to be compared. Aside from an objective diagnosis, beneficial real-world applications of subject classification include: the pre-symptomatic diagnosis of AD, or the prediction of which subjects will progress from MCI to AD.

Statistical predictors such as classifiers and regressors take as input some feature data $\mathbf{d}$, and output an estimated outcome variable $\hat{o}$. Therefore, for a given class of statistical predictor, h, $\hat{o} = h(\mathbf{d})$.

In this thesis, only the class of supervised learners are considered. Supervised learning requires a "training set" of N data items that are labelled with a known outcome variable, o. Here, a training set is defined as:
$\mathcal{L} = \{(o_n, \mathbf{d}_n), n = 1, 2, ..., N\}$, where n indexes the subjects in the training set. Statistical predictors estimate the relationship between $\mathbf{d}$ and o from $\mathcal{L}$ in a process known as training. The trained predictive model can provide an estimate for a test subject i, based on $\mathcal{L}$, $\hat{o}_i = h(\mathbf{d}_i, \mathcal{L})$.

There are many classifier variants, and a full review is outside the scope of this thesis. Two commonly used classifiers are used in this thesis: support vector machines and naïve Bayes. They are described in the following sections.

Support Vector Machines

Support vector machines [40] (SVMs) are maximum margin classifiers, thus estimate the separating hyperplane that discriminates between training examples of different classes that are maximally distant from the nearest training example. This approach leads to a classification model that generalises well, and accurately predicts the class of unseen data. The hyperplane estimation uses only a small amount of the training examples that are near the hyperplane. These training examples are called support vectors.

The hyperplane is defined in a classification feature space, $\mathbf{z}$, as:

$$\mathbf{w}^T\mathbf{z} + b = 0 \tag{2.30}$$

where the weights of the hyperplane $\mathbf{w}$ are a linear combination of support vectors:

$$\mathbf{w} = \sum_{n}^{Ns} \alpha_n \phi(\mathbf{d}_{sv})_n \tag{2.31}$$

where ϕ maps between the original data space of $\mathbf{d}$ and $\mathbf{z}$. n indexes through the N_s support vectors $\mathbf{d}_{sv} \subseteq \mathcal{L}$. b is a bias term of the hyperplane.

This leads to a classification function for an unseen testing example, $\mathbf{d}_{test}$:

$$\hat{o} = \text{sign}\left(\sum_{n}^{Ns} \alpha_n \phi(\mathbf{d}_{train})_n \cdot \phi(\mathbf{d}_{test}) + b\right) \tag{2.32}$$

where $\cdot$ refers to the dot product between the vectors containing the feature data. The comparison of features in a high dimension space can be made computationally tractable through the use of the so-called kernel trick. This implicitly maps each data example into a higher dimensional feature space. These kernels allow the dot product between two examples to be efficiently calculated in a higher dimensional space, based on the original, low-dimensional data. This provides classification decision surfaces that are non-linear with respect to the original data space. Kernels that induce a high dimensional feature mapping will require parameters to be selected. For example, the width of the radial basis function, or the degree of the polynomial.

The most common formulation of SVMs contain an additional parameter, which is not inferred as part of the optimisation. This parameter C, permits a flexible penalty system to account for the misclassification of training examples [40]. Such a framework leads to a soft-margin classifier, which means that the SVM can be robustly trained on data that is not linearly separable in the feature space, $\mathbf{z}$. The use of a soft-margin leads to the hyperplane being optimised with respect to a different cost function, where there is no penalty for correctly classified examples that are outside of the margin. The misclassification cost for correctly classified support vectors that are inside the margin is: $0 < \xi_n \leq 1$. For misclassified examples, the classification cost is $\xi \geq 1$, and increases linearly with distance from the separating hyperplane. The function being minimised in this formulation is:

$$C\sum_{n}^{Ns} \xi_n + \frac{1}{2}||\mathbf{w}||^2 \tag{2.33}$$

Smaller values of C leads to a smaller margin, and therefore less support vectors are required, although there will be greater misclassification error of the training set. Conversely, higher values of C result in more accurate discrimination of the training set, but wider margins, which leads to more support vectors. This makes it more likely that the classifier will over-fit the training data, and not perform as well on unseen data. C needs to be selected to provide an optimal trade-off.

A principled manner in which it can be chosen is described in the section below entitled parameter selection.

Support vector machines are often used in classification problems in neuroimaging due to their efficiency and ability to generalise to previously unseen data. In this thesis, the LibSVM implementation of SVMs is used [33].

Gaussian Naïve Bayes

Gaussian naïve Bayes is an effective and efficient approach to data classification [204]. It has an additional benefit of not requiring any parameter tuning. The naïve Bayesian classifier estimates the probability of an example belonging to a class based on the set of data features, in this case voxels. From Bayes' rule it can be seen that:

$$p(o|\mathbf{d}) = \frac{p(o)p(\mathbf{d}|o)}{p(\mathbf{d})} \tag{2.34}$$

However, the joint probability density of the data $\mathbf{d}$ is likely to be difficult to calculate, especially for a large number of voxels. Therefore a naïve assumption of independence between the variables in $\mathbf{d}$ is assumed, namely:

$$p(o|\mathbf{d}) = \frac{1}{Z}p(o)\prod_{j}^{N_j} p(\mathbf{d}_j|o) \tag{2.35}$$

where Z is a normalisation constant dependent on $\mathbf{d}$, j indexes the features in the feature vector, N_j is the total count of features and $p(o)$ gives the prior probability of observing a particular class o. $p(\mathbf{d}_j|o)$ is assumed to follow a Gaussian distribution: $p(\mathbf{d}_j|o) = \mathcal{N}(\mu_j, \sigma_j)$. The parameters of each Gaussian are fitted using maximum likelihood based on the training set, $\mathcal{L}$. Once the model parameters have been learned from the training data, the probability of a new data example belonging to a class can be can be estimated using equation 2.35. The class attributed the highest probability given the data, is selected as the most likely class. Despite the assumption of independence between voxels being incorrect, naïve Bayes has been demonstrated to provide accurate classification. Although, the probabilistic estimates of the class of an example are typically unreliable.

Parameter Selection

The majority of approaches to statistical classification require the selection of hyper-parameters, including the SVMs described previously. The optimal pa-

rameters need to be derived in such a manner as to preserve the fairness of the test. This requires that the test data, along with its group labelling, is not involved in parameter selection. A principled mechanism for inferring an optimal set of parameters that should generalise to unseen data, given a labelled training set of data $\mathcal{L}$, is cross validation. Cross validation estimates the generalisation accuracy of a statistical prediction tool, by training and testing on different subsets of the training data. A common approach is to divide the training set into k approximately equally sized, and class balanced "folds". The performance of the model, under a given parametrisation, is then assessed by training the model on $k-1$ folds of data and validating on the left out fold. This is repeated for each fold, and averaged to reduce the variability in the estimation. Using a larger number of folds provides more stable estimates of performance as the variability between training folds is reduced, but results in greater computational expense.

Feature Selection

To allow accurate and efficient classification or regression, it may be necessary to reduce the dimensionality of the set of feature data. MR images of the brain may contain several million voxels, some of which will be highly correlated. Inferring on a relationship between each of these voxels and an outcome variable may be slow to calculate and erroneous. Therefore, several approaches have been proposed to reduce the dimensionality of the data. The simplest approach is to use a region of interest (ROI) analysis. By manually selecting a range of interesting voxels, some of the difficulties in inference can be reduced. However, this requires that a suitable ROI is known in advance. An alternative approach, is to select voxels based on consistent differences between groups. These group differences can be evaluated in a voxelwise fashion using classical statistical tests, such as t-tests.

Alternative approaches to reducing the number of features are data decomposition approaches. Examples of such methods include principal and independent component analysis that represent a set of data through a linear basis set of image features. These methods may follow certain assumptions, such as Gaussianity of the variables, but could be used to provide an efficient compression of the data.

2.3.5 Summary

This section has introduced two types of morphological features that can be used to describe differences in brain anatomy, voxel-, and tensor-based morphometry.

Additionally, the role of registration in each of these feature types is discussed. Approaches for the generation of representative atlas images are introduced, and two methods for the statistical prediction of disease status based on image data are described.

The next chapter introduces a generative model for non-rigid registration that is inferred upon using variational Bayes. This framework allows the inference of the level of regularisation and data fidelity, and provides estimates of the uncertainty in registration.

Chapter 3

Probabilistic Non-Rigid Registration with Inferred Regularisation

3.1 Introduction

This chapter introduces a principled approach to the data-driven inference of the parameters that control the level of regularisation and data fidelity in non-rigid image registration models. This adaptive approach to medical image registration allows flexible treatment of different data and hierarchical registration schemes without necessitating any manual tuning of the parameters. A further benefit is that it provides an estimated measure of registration uncertainty, which arises naturally from the probabilistic framework. This work has been published [169], with an earlier version in [163].

3.1.1 Motivation

As was highlighted in the previous chapter, accurate registration is a vital step in performing population, or longitudinal imaging studies. A limitation that exists in almost all registration methods is that the user is required to select some parameters to ensure that the inferred mappings between images are both accurate and plausible. The majority of these parameters can be selected for

a wide range of subjects, e.g. degrees of freedom of the transformation, image sub-sampling and smoothing. However, the vast majority of non-rigid registration approaches require the definition of an additional parameter, or parameters, which are likely to need manual tuning for different datasets. These are the parameters that control the trade-off between image fidelity, how much the image data is trusted, and regularisation, which aims to keep the inferred mapping simple and smooth.

The relative trade-off of these parameters is key for inferring a reasonable mapping between subjects. This trade-off describes the expected transformation complexity, as defined by the regularisation model, for a given measurement of image similarity. Where too low an emphasis is placed on regularisation, the expected transformation complexity will be greater than is required. This may lead to spurious matching of voxels based on irrelevant intensity differences, which results in unnecessarily large and complex mappings. Conversely, if the influence of regularisation is too strong, the inferred mapping is likely to be inaccurate. The level of regularisation or data fidelity may require tuning due to variability in the signal-to-noise (SNR), or contrast, of the images being registered. Where images with poor SNR, or different contrasts, are being registered, the optimisation may require stronger constraints to resist attempting to register image noise, as opposed to structure. Furthermore, subjects will be anatomically more similar to some, than others. This leads to an expectation that registration between different subjects will require a variable level of transformation complexity. Therefore, a "one size fits all" approach to penalising the complexity in registration will allow for the possibility of over-, or under-constraining the mapping in some circumstances.

3.1.2 Previous Approaches to Parameter Selection

Manual

Regularisation parameter values have traditionally been selected using a trial and improvement strategy [156][5][7]. Here, a user finds an appropriate set of parameters that provide qualitatively reasonable results over a specific set of data. As hierarchical registration schemes are commonly utilised in non-rigid registration, such an approach may require a user to hand-tune several parameters. Manual attempts to parameter selection are subjective and time consuming.

Cross-validation

The simplest objective approach to parameter tuning is through cross-validation. Cross-validation allows the assessment of the quantitative effects, according to a given metric, of a particular set of parameters on a training set of data. The parameter set that produces the optimal score on the training set are selected for use in subsequent registrations. This framework can be used to find the most effective cost-function [201], which could vary spatially [202], for a particular task. Such a procedure has been demonstrated using the localisation accuracy of cortical folds as an objective metric [202]. This procedure requires manually labelled data to train on that is representative of the testing data of interest

However it is derived, using a fixed level of regularisation on a set of images makes the assumption that all data require a similar level of regularisation, whereas the optimal level of regularisation will have a dependence on the data presented to it.

Probabilistic Inference

There has been some work in the field of medical image registration to facilitate the probabilistic inference of regularisation. Van Leemput [183] uses a generative model to describe the formation of a labelled probabilistic image atlas as a result of groupwise registration. This approach uses a finite element transformation model, regularised using a Markov random field prior. The strength of the prior describes the flexibility of all the mappings between a template and a set of subjects. The value that provides the optimal segmentation is inferred from the data. He does not provide quantitative results on 3D data due to the difficulties of integrating over the distribution of possible segmentations.

More closely related to the proposed method is the work of Risholm et al. In their work, the elastic material parameters in intra-subject brain registration are estimated from the data [146]. Their approach uses a generative model for registration with the two Lamé parameters, that govern the elastic energy prior, described as random variables. The elastic prior may vary over space, or across spatial compartments. This model is inferred upon using Markov chain Monte Carlo, which does not require assumptions to be made about the distribution of the posterior. This has a major disadvantage in that the computational complexity of numerically integrating over the model parameters constrains this approach to only being feasible for low resolution registration.

3.1.3 Proposed Solution

This chapter proposes a generative model for non-rigid registration, which is inferred upon using an approximate full Bayesian inference technique. This allows for the data-driven and computationally tractable inference of the parameters that controls the level of regularisation and data fidelity.

3.2 A Generative Model of Image Registration

The process of image registration can be described probabilistically through the use of a generative model. As described in section 2.1.4, the majority of generative models for registration use a Gaussian noise model, which is equivalent to a sum of squared differences (SSD) cost function. This cost function is appropriate for single modal MR brain registration when coupled with an intensity mapping to model the difference in contrast between the two images and any bias fields [5]. Therefore, a Gaussian likelihood is chosen due to its simple and efficient formulation.

As a brief recap of section 2.1.1, a generative model for image registration assumes that the target image data, $\mathbf{y}$, can be generated from a source image, $\mathbf{x}$, when it is deformed according to some transformation model. Here, $\mathbf{t}(\mathbf{x},\mathbf{w})$ is the transformed source image, where $\mathbf{w}$ parametrises the transformation. In all equations, column vectorised forms of $\mathbf{y}$, $\mathbf{t}(\mathbf{x}, \mathbf{w})$ and $\mathbf{w}$ are used.

The specific form of the transformation model used in $\mathbf{t}$ for this implementation is a B-spline free-form deformation (FFD), as described in section 2.1.1. As such, $\mathbf{w}$ represents the set of B-spline control point displacements in each direction. The B-spline FFD model was chosen to demonstrate this approach as it requires substantially fewer parameters to optimise than a time-varying [22], or stationary, [7][185] velocity field approach, whilst providing a mechanism for the efficient and sparse calculation of the covariance between transformation parameters. Although stationary velocity fields [7], or initial velocities in geodesic shooting [14], permit the rapid calculation of transformation parameter covariance, the number of parameters renders these options computationally expensive. Velocity fields could potentially be parameterised by basis functions, such as in Modat et al. [126], but the covariance between the transformation parameters will likely be very difficult to estimate.

The generative model includes a noise term, which represents the residual difference in intensities between the target, and transformed source images. These

intensity differences will be caused by a combination of: noise in the image formation process, image structures that cannot be matched using the given transformation model and resolvable image misalignment. The noise term is assumed to be independent and identically distributed (i.i.d.) across image voxels, and follows a normal distribution, $\mathcal{N}$:

$$\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \phi^{-1}\mathbf{I}) \tag{3.1}$$

where $\mathbf{I}$ represents the identity matrix, and ϕ is a global precision (inverse variance) of the noise across the image.

Including this noise term, a generative model for registration can be formulated as:

$$\mathbf{y} = \mathbf{t}(\mathbf{x}, \mathbf{w}) + \mathbf{e} \tag{3.2}$$

As the noise is assumed to be i.i.d Gaussian the likelihood of any given voxel in the target image, indexed by i, is given by:

$$P(\mathbf{y}_i|\mathbf{x}, \mathbf{w}, \phi) = \left(\frac{\phi}{2\pi}\right)^{\frac{1}{2}} \exp^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{t}(\mathbf{x},\mathbf{w})_i)\phi(\mathbf{y}_i - \mathbf{t}(\mathbf{x},\mathbf{w})_i)} \tag{3.3}$$

From equation (3.3) the log-likelihood of the target image data, $\mathbf{y}$, given the set of model parameters can be defined as:

$$\log p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \phi) = \frac{N_v}{2} \log \frac{\phi}{2\pi} - \frac{\phi}{2}(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w}))^{\mathsf{T}}(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})) + \kappa \tag{3.4}$$

where κ contains all terms that are constant with respect to the model parameters. N_v is the number of voxels accounted for by the model. The choice was made to discount any voxels that contains background information in both images from the model. This is because these voxels do not contain relevant information regarding the matching of anatomy. Therefore, their inclusion provides a discordant contribution to the estimation of the i.i.d. noise term, $\mathbf{e}$, for the task of brain registration.

To regularise the registration, prior distributions over the parameters are included. This provides a full probabilistic model, which is graphically described in Figure 3.1.

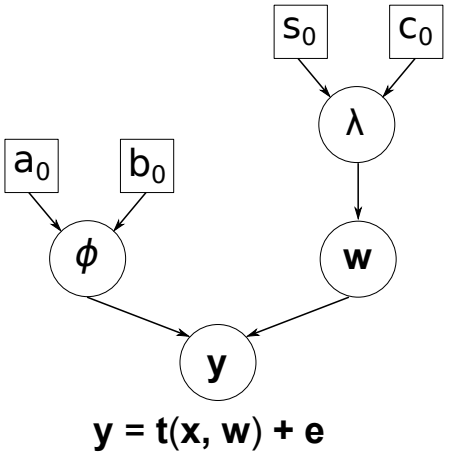

Figure 3.1: A graphical description of the probabilistic dependencies of the registration model parameters. The variables in square boxes are constants, and the variables in circles are random variables. The image $\mathbf{y}$ is noisily generated from the transformed source image $\mathbf{t}(\mathbf{x}, \mathbf{w})$, where $\mathbf{x}$ is the source image, and the transformation is parametrised by $\mathbf{w}$. The prior on $\mathbf{w}$ is given in equation 3.5 and is parametrised by λ. The Gaussian noise, $\mathbf{e}$, of this model has a single parameter, ϕ.

3.2.1 Priors

Transformation parameters

As described in section 2.1.3, the set of transformation parameters $\mathbf{w}$ needs to be spatially regularised. Regularisation preserves the topology of the source image, by enforcing spatial covariance in neighbouring transformation parameters. It also penalises the registration complexity such that complex transformations are not inferred, without being supported by sufficient image information. Transformation complexity is a measured by the deviation of $\mathbf{w}$ from the spatial prior. In an elastic approach to regularisation, the spatial prior has a mean of the identity transformation, which is where $|\mathbf{w}| = 0$. It also has some covariance structure, which describes smooth transformations. Therefore, greater deviation from the spatial prior means larger, and less smooth transformations.

Regularisation is incorporated into this probabilistic framework by assigning an appropriate prior distribution to $\mathbf{w}$. $p(\mathbf{w})$ is modelled using a multivariate Normal distribution:

$$\begin{aligned} p(\mathbf{w}|\lambda) &= \mathcal{N}(\mathbf{w}; \mathbf{0}, (\lambda\mathbf{\Lambda})^{-1}) \\ &= \frac{|\lambda\mathbf{\Lambda}|^{\frac{1}{2}}}{(2\pi)^{\frac{N_C}{2}}} \exp^{-\frac{1}{2}\mathbf{w}^{\mathsf{T}}(\lambda\mathbf{\Lambda})\mathbf{w}} \end{aligned} \tag{3.5}$$

The form of the prior knowledge of $\mathbf{w}$ is described in equation (3.5). Here, $\mathbf{\Lambda}$ encodes bending energy regularisation, although other models could be chosen, as an $N_c \times N_c$ spatial kernel matrix. Λ is specified in the form of a precision (inverse covariance) matrix. This representation is preferred to a covariance matrix as it has a sparse form, unlike the covariance matrix. N_c is the count of all the transformation parameters that have any effect on the likelihood term. In the case of the B-spline FFD transformation model, this is the number of control points that have any effect on the foreground image data, in all three deformation directions. λ is a scalar spatial precision parameter that controls the level of regularisation. λ is modelled as a random unknown variable, therefore it can be determined adaptively from the data resulting in an automated approach to regularisation. In the case where λ is given a fixed value, the approach will correspond to other generative approaches to registration, such as DARTEL [7] or FNIRT [5]. The novelty of the proposed approach lies in the *inference* of λ based on the data.

Spatial precision

As the spatial precision parameter, λ, is probabilistically modelled, a prior distribution on λ must be specified. The prior on λ is modelled using a Gamma (Ga) distribution:

$$\begin{aligned} P(\lambda) &= Ga(\lambda; s_0, c_0) \\ &= \lambda^{c_0-1} \frac{\exp^{-\frac{\lambda}{s_0}}}{\Gamma(c_0) s_0^{c_0}} \end{aligned} \tag{3.6}$$

Equation 3.6 shows the definition of the prior over λ, with initial scale, s_0, and shape, c_0, parameters. A Gamma distribution can be used to set an uninformative prior over the possible values of λ, with the distribution parameters set to $s_0 = 10^{10}$, $c_0 = 10^{-10}$.

Noise precision

In order to evaluate the optimal value of λ, the level of noise in model fit also needs to be accurately estimated. This is because of the inherent trade-off between regularisation and maximising the likelihood. Therefore, ϕ also needs to be inferred during the registration. The model noise is assumed to follow a Gaussian distribution, that is independent and identically distributed across voxels. This

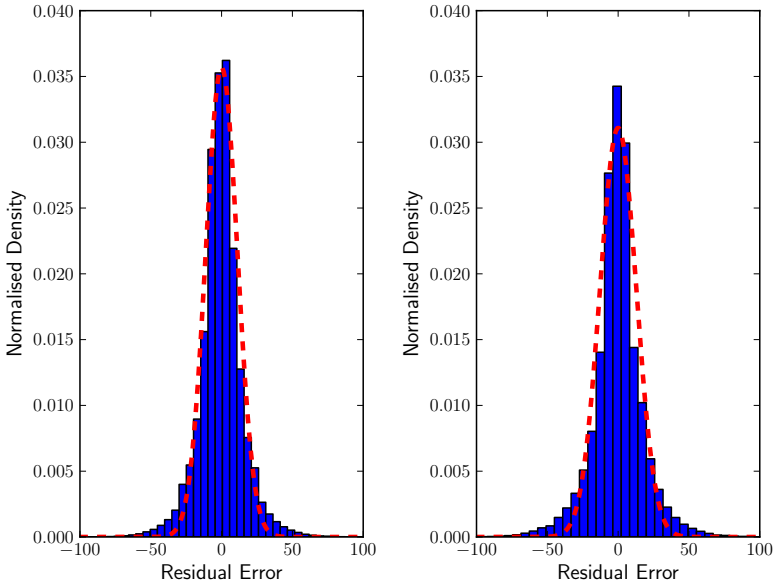

Figure 3.2: Two example histograms of the un-smoothed residual image, $\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})$, after fitting the registration model parameters for two pairs of high resolution (1 mm^3), high SNR, T1 weighted MR images of different human brains taken from the NIREP database [35]. The dashed overlaid red line shows the estimated i.i.d. Gaussian noise precision. In both of these cases it can be seen that the centre of the residual distribution is well modelled by the Gaussian. However, the residual distribution does have heavier image tails, resulting in a positive kurtosis of 3.4, for the left and 4.7 for the right histogram.

Gaussian has zero mean and variance ϕ^{-1}. The prior on ϕ is modelled as being Gamma distributed:

$$\begin{aligned} P(\phi) &= Ga(\phi; a_0, b_0) \\ &= \phi^{b_0 - 1} \frac{\exp^{-\frac{\phi}{a_0}}}{\Gamma(b_0) a_0^{b_0}} \end{aligned} \tag{3.7}$$

where a_0 and b_0 are the initial scale and shape prior hyper-parameter estimates of the distribution. These are again chosen to give a non-informative prior distribution with $a_0 = 10^{10}, b_0 = 10^{-10}$.

3.2.2 Noise Model

As described in section 3.2, the noise in model fit is approximated to be zero mean, independently and identically distributed Gaussian noise. The appropriateness of

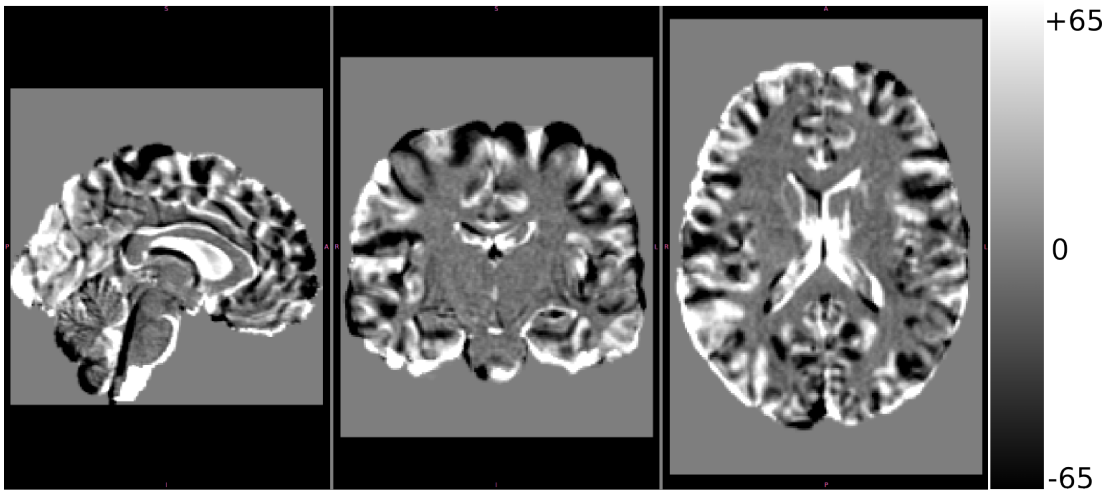

Figure 3.3: An initial residual image, $\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})$, from two high resolution (1 mm^3) T1 MR images of different human brains taken from the NIREP database [35]. These images have almost identical contrast, and high SNR. The images are affinely aligned using FLIRT [98]. As can be seen, there are large clusters of residuals in the model fit, particularly around the cortical regions and the ventricles. The residuals are clearly highly spatially correlated due to image misalignment. As the alignment improves, the noise will become less correlated.

this approximation for inter-subject registration is illustrated in Figure 3.2 using histograms of the residual image ($\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})$) from fitted registration models, overlaid with the inferred Gaussian noise distribution. Although the centre of the distribution is well approximated as a Gaussian, the distribution has positive kurtosis, and is therefore not Gaussian distributed. In inter-subject brain registration misalignment of anatomy, rather than image formation noise, will be a large cause of error in model fit. These "outlier" voxels introduce heavier tails to the residual image distribution. This could potentially be address in future work by the use of a mixture of Gaussians [194]. Alternatively, as the misaligned structures are likely to be spatially localised, the correctness of the noise model could be improved by describing the residual as locally Gaussian, and this is addressed in the Chapter 4.

The tails of the distribution in real data will be affected by the variability of structures within the field of view, and the variability in contrast of some non-brain structures. These outlier voxels can be removed through the extraction of the brain from the image using a suitable tool, such as the brain extraction tool (BET) [170].

Covariance in the Residual Image

Unfortunately, the assumption of spatial independence in the registration noise model is largely incorrect. There are two primary sources of spatial noise covariance to consider; firstly, image data is often pre-smoothed using a Gaussian filter to increase the SNR of the data and preserve features of a specific scale. This is performed with different full-width at half-maximum (FWHM) values at different levels of the hierarchical registration scheme. Image smoothing introduces additional covariance in the image data, and therefore also in the noise. Secondly, and more importantly, the misalignment of tissue during the registration process introduces spatial correlation in the residual image. This is due to regions of tissue types often being spatially contiguous. An example difference image after affine registration is given in Figure 3.3, which illustrates the spatial correlation of the model fit due to misalignment. The spatial covariance in $\mathbf{e}$ needs to be compensated for, to avoid over-emphasising the noise precision, and therefore the relative importance of the likelihood to the spatial prior.

The most direct method to compensate for the spatial covariance of the residual noise, $\mathbf{e}$, is to model spatially smooth noise using a Gaussian process. This would allow an adaptive determination of the noise covariance. Unfortunately, this approach is computationally very demanding. An approximate solution, with minimal computational overhead is available. This is based on estimating the number of RESELS (RESolution ELementS). The number of RESELS is an approximation of the number of independent signals in the data. If this residual noise is assumed as having been smoothed using a Gaussian kernel, the degrees of freedom of the unsmoothed image can be approximated using Gaussian random field theory [198]. All terms that sum over voxels are weighted by the ratio of the degrees of freedom in the image to the number of voxels in the overlap N_v. This is equivalent to decimating the data, a process that reduces the number of data samples to remove redundancy. However, as decimating the data requires removing voxels that may still contain valuable information, a weighting term is used instead, providing a virtual decimation (VD) [73]. The VD weighting factor, α, can be calculated using equation (3.8) as given in [198]:

$$\alpha = \frac{\text{RESELS}}{N_v}(4\log(2)/\pi)^{D/2} = \left(\frac{0.9394}{\text{FWHM}_x}\right)\left(\frac{0.9394}{\text{FWHM}_y}\right)\left(\frac{0.9394}{\text{FWHM}_z}\right) \quad (3.8)$$

where $\text{FWHM}_{\{x,y,z\}}$ is the full width at half maximum of the equivalent Gaussian smoothing kernel in each direction. The smoothing kernel's FWHM is estimated

by assessing the correlation between adjacent voxels in each direction:

$$(\text{FWHM}_l)^2 = -2\log(2)/\log(\text{corr}_l) \tag{3.9}$$

where corr_l is the correlation between adjacent voxels in the direction l, $l \in \{x, y, z\}$. This adjustment weights all terms that sum over voxels such that only the approximate number of "independent" noise observations is considered. Although this approach is non-Bayesian, it is still determined from the data, and therefore fits the adaptive framework that is being considered. This approach has also been used in the registration approach of Ashburner and Friston [10].

3.3 Model Inference

The previous section described the probabilistic generative model for non-rigid registration between two MRI images of the human brain. This generative model, described in Figure 3.1, contains a set of unknown parameters and hyper-parameters, these are modelled as random variables. These random variables provide a probabilistic description of the transformation parameters, $\mathbf{w}$, the strength of the prior distribution on $\mathbf{w}$, λ, as well as describing the error in model fit, ϕ. As a hierarchical prior model is specified, a full Bayesian inference approach is required. Variational Bayes, as described in section 2.2.4, is used to provide tractable full Bayesian inference.

3.3.1 Mean Field Approximation

To allow tractable inference with VB, the mean-field approximation was applied. For this case the transformation, noise and regularisation parameter distributions are factorised in the approximate posterior:

$$p(\mathbf{w}, \phi, \lambda|\mathbf{y}) \sim q(\mathbf{w}, \phi, \lambda) = q(\mathbf{w})q(\phi)q(\lambda) \tag{3.10}$$

3.3.2 Variational Free Energy

Using the VB framework, analytic updates were derived for the approximate posterior distributions of the transformation, regularisation and noise parameters, which seek to maximise the negative variational free energy $\mathcal{F}$. In the case of this

registration model using the approximation in equation 3.10, $\mathcal{F}$ can be written as the sum of four terms:

$$\mathcal{F} = \mathcal{L}_{av} - D_{KL}(q(\mathbf{w})\|p(\mathbf{w})) - D_{KL}(q(\lambda)\|P(\lambda)) - D_{KL}(q(\phi)\|P(\phi)) \quad (3.11)$$

where $\mathcal{L}_{av}$ is the expectation of the log-likelihood given in equation 3.4, with respect to both the Gamma distribution on ϕ and the Gaussian noise model. $D_{KL}(q\|p)$ is the Kullback-Leibler (KL) divergence [108] between the approximate posterior, q, and the prior, p, for each of the groups of model parameters. The KL terms penalise over-confidence in the parameter estimates and deviation from the prior distribution. The full definition of $\mathcal{F}$ for this registration model is provided in Appendix A. The derivation of the updates presented below is provided in Appendix B.

3.3.3 Inference on Transformation Parameters

The approximate posterior distribution of the transformation parameters, $\mathbf{w}$, follows a multivariate Normal distribution:

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Upsilon}^{-1}) \quad (3.12)$$

where $\boldsymbol{\mu}$ is an $N_c \times 1$ vector containing the posterior mean, and $\boldsymbol{\Upsilon}^{-1}$ is the $N_c \times N_c$ posterior covariance matrix of $\mathbf{w}$.

As VB is not tractable for arbitrary non-linear forward models, a first order Taylor series approximation is used to provide a linear approximation of the transformed image, with respect to $\mathbf{w}$:

$$\mathbf{t}(\mathbf{x}, \mathbf{w}) \approx \mathbf{t}(\mathbf{x}, \boldsymbol{\mu}) + \mathbf{J}(\mathbf{w} - \boldsymbol{\mu}) \quad (3.13)$$

where $\mathbf{J}$ is the $N_c \times N_v$ matrix of first order partial derivatives of the estimated transformed image, $\mathbf{t}(\mathbf{x},\mathbf{w})$, with respect to $\mathbf{w}$, centred on $\boldsymbol{\mu}$.

Through VB, analytic updates can be derived for $\boldsymbol{\mu}$ and $\boldsymbol{\Upsilon}$:

$$\boldsymbol{\Upsilon} = (\alpha\bar{\phi}\mathbf{J}^T\mathbf{J} + \bar{\lambda}\boldsymbol{\Lambda}) \quad (3.14)$$

$$\boldsymbol{\Upsilon}\boldsymbol{\mu}_{new} = \left[\alpha\bar{\phi}\mathbf{J}^T(\mathbf{J}\boldsymbol{\mu}_{old} + \mathbf{k})\right] \quad (3.15)$$

where $\mathbf{k}$ is a length $N_v \times 1$ vector representing the residual image $\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})$. The subscripts *new*, and *old* of $\boldsymbol{\mu}$ give the previous, and new estimate of $\boldsymbol{\mu}$ respectively. $\bar{\lambda}$ is the expectation of the posterior spatial precision distribution, and $\bar{\phi}$ is the expectation of the estimated noise precision. As both $\boldsymbol{\mu}_{new}$ and $\mathbf{J}$ depend on the previous parameters, $\boldsymbol{\mu}_{old}$, these updates need to be applied iteratively until the convergence criteria is met.

These updates are equivalent to those in Gauss-Newton as described in section 2.1.6, and for fixed values of $\bar{\phi}$ and $\bar{\lambda}$ this method is equivalent to standard non-linear least-squares approaches to registration such as FNIRT [5].

3.3.4 Inference on Regularisation Parameters

The variational Bayesian methodology can be used to provide updates on the posterior distribution of the regularisation control parameter, λ. The approximate posterior distribution of λ is Gamma distributed, $q(\lambda) = Ga(\lambda; s, c)$. The distribution parameter updates are as follows:

$$c = c_0 + \frac{N_c}{2} \tag{3.16}$$

$$\frac{1}{s} = \frac{1}{s_0} + \frac{1}{2}\left(Tr(\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Lambda}) + \boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\right) \tag{3.17}$$

where Tr refers to the matrix trace operation. The expectation of the approximate posterior distribution over λ, is given as $\bar{\lambda} = sc$.

3.3.5 Inference on Noise Parameters

The approximate posterior noise parameter distribution, ϕ, is Gamma distributed, $q(\phi) = Ga(\phi; a, b)$, with distribution parameter updates given by:

$$b = b_0 + \frac{N_v\alpha}{2} \tag{3.18}$$

$$\frac{1}{a} = \frac{1}{a_0} + \frac{1}{2}\alpha(\mathbf{k}^T\mathbf{k} + Tr(\boldsymbol{\Upsilon}^{-1}\mathbf{J}^T\mathbf{J})) \tag{3.19}$$

As described in section 3.2.2, N_v is scaled by the virtual decimation factor, α, such that it measures the number of independent noise voxels. This is required to prevent over-emphasising ϕ. The expectation of the approximate posterior distribution over ϕ, is given as: $\bar{\phi} = ab$.

3.3.6 Informative Prior Distributions on λ and ϕ

The model described in this Chapter assumes an uninformative prior distribution of λ and ϕ. This allows the inference framework to select the parameter values in an entirely data driven manner. A limitation is that this may be more susceptible to local minima, as there is no prior constraint that λ and ϕ should be within certain plausible ranges. A solution to this is the use of an informative prior distribution, as derived from the registrations of a range of subjects. An informative Gamma distribution for λ and ϕ could be inferred by maximum likelihood fitting of a population of inferred parameters.

Such an approach would ideally be performed with many registrations in parallel, where $P(\lambda)$ and $P(\phi)$ are re-calculated every iteration according to the values from all of the registrations. Unfortunately, such an approach would require communication between many simultaneous registrations, which would be complicated to implement in practise. An alternative approach is to perform the registration using a representative training set, or all of the data with an uninformative prior. An informative prior can then be derived from the population of converged parameter values at each level of the hierarchical registration scheme. The population can then be registered using the derived informative priors.

3.4 Implementation

3.4.1 Approximations

From a practical perspective, the update terms required for both the noise (equation 3.19), and spatial (equation 3.17), precisions contain terms that are computationally infeasible to calculate. Specifically, the difficult term to compute is the inverse of the posterior precision matrix of the transformation parameters, $\boldsymbol{\Upsilon}^{-1}$. As $\boldsymbol{\Upsilon}$ is a large sparse matrix, calculating the inverse is computationally very intensive, and would require a very large amount of memory. Therefore, only for the updates in equations 3.19 and 3.17, an approximation to the posterior covariance matrix is made, which assumes that the control point at each location is independent of its neighbours and only has cross-directional covariance. This allows a sparse inverse approximation that can be rapidly calculated and provides a sufficiently accurate estimation of ϕ and λ. The accuracy of this approximation is described in Appendix C.

For calculating the update on $\boldsymbol{\mu}$ in equation 3.15, it is not necessary for $\boldsymbol{\Upsilon}^{-1}$ to be explicitly calculated. Instead, approximate methods can be used on the full precision matrix $\boldsymbol{\Upsilon}$ to find $\boldsymbol{\mu}_{new}$; in this implementation a conjugate gradient method was used.

Although the updates seek to maximise the negative variational free energy, $\mathcal{F}$, this is not calculated in full for measuring the convergence of each update. This is because of the large computational expense of calculating the log determinant of very large matrices. Instead, equation 3.20 is calculated to measure convergence after each update for $\boldsymbol{\mu}$.

$$\mathcal{C} = \alpha\bar{\phi}\mathbf{k}^T\mathbf{k} + \bar{\lambda}\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu} \tag{3.20}$$

3.4.2 Software

FNIRT

The 3D implementation of this registration model was incorporated into the FMRIB Non-linear Image Registration Tool (FNIRT) [5]. FNIRT is a non-linear least squares (NLLS) implementation of a FFD B-spline model as popularised by [156]. FNIRT uses MAP inference with Gauss-Newton optimisation. It is regularised through a fixed-parameter bending energy prior, with a simple noise model that depends on the SSD of the image residual. The FNIRT λ parameter encodes the trade-off of data fidelity and regularisation into a single parameter. FNIRT uses a non-linear intensity mapping to account for differences in image contrast. The parameters of this mapping are estimated concurrently with the transformation parameters. FNIRT was chosen as a basis for implementation of the described model due to its formulation as a generative model, solved in a NLLS framework. FNIRT also has an efficient mechanism for calculating the Hessian matrix, which would otherwise be computationally expensive. However, it must be stressed that this work describes a generic framework and is not restricted to application in FNIRT. It could be implemented in any other generative model regularised through an elastic prior on the transformation parameters, including the diffeomorphic approaches of a stationary velocity field [7] or using geodesic shooting [14], which may have some advantages in mapping larger deformations.

Algorithm 1 Pseudo-code description of the VB registration algorithm:

$\boldsymbol{\mu} = 0$
Run 1st hierarchical level using fixed trade-off using the FNIRT default.
for $i = 2$ to number of hierarchical levels **do**
  if Informative prior **then**
    Set $a = a_0$,$b = b_0$,$s = s_0$,$c = c_0$
  else
    Set a_0,b_0,s_0,c_0 as uninformative
  end if
  Smooth and sub-sample images according to the hierarchical scheme
  α=VD(**y**-**t**(**x**,**w**)) // eq. (3.8)
  Refine $\boldsymbol{\mu}$ to new resolution level
  Calculate **J** // Re-linearise first order derivatives of $\mathbf{t}(\mathbf{x}, \boldsymbol{\mu})$
  while a, b, s, c not converged **do**
    Update $\boldsymbol{\Upsilon}$ // eq. (3.14)
    Update a, b, s, c // eqs. 3.16 to 3.19
  end while
  while Calculate$\mathcal{C}() > \mathcal{C}_{old}$ **do**
    $\mathcal{C}_{old}$ = Calculate$\mathcal{C}()$ // eq. 3.20
    Calculate **J**
    Update $\boldsymbol{\Upsilon}$
    Update $\boldsymbol{\mu}_{new}$ // eq. 3.15
    if Calculate$\mathcal{C}() > \mathcal{C}_{old}$ **then**
      $\boldsymbol{\mu} = \boldsymbol{\mu}_{new}$
      Update a, b, s, c
    end if
  end while
end for

Psuedocode

An algorithmic summary of the proposed method is presented in Algorithm 1. Unlike in the previously published work [169], the coarsest level of the hierarchical registration scheme uses the default FNIRT trade-off. This is because of the difficulties in registering data that requires a complicated intensity mapping. When using the proposed algorithm for the coarsest level of the hierarchical registration scheme, the intensity mapping is not initially well estimated, resulting in a high level of regularisation. This under-estimates the larger scale warps, increasing the chances of being stuck in a weak local minimum. By estimating the coarsest level with a fixed trade-off, a reasonable estimate of the intensity mapping and large scale warps can be found. For subsequent levels of the hierarchical registration scheme, the model hyper-parameters are initially optimised based on the current transformation to aid faster convergence.

Hierarchical registration level	Image sub-sampling factor	Control point spacing (mm)	Image smoothing FWHM (mm)	Fixed FNIRT λ
1A	8	40	8	300
1B	8	40	6	150
2A	4	20	5	100
2B	4	20	4.5	50
3A	2	10	3	40
3B	2	10	2	30
4A	1	5	2	40/20
4B	1	5	1.5	30/10

Table 3.1: Hierarchical registration scheme used by FNIRT, extended to have a final 5mm control point spacing, with either a lower or higher regularisation at the finest two levels.

3.4.3 Hierarchical Registration Scheme

FNIRT uses both a multi-resolution, and a multi-level approach to image registration, as described in section 2.1.5. In this thesis, to allow comparison with FNIRT, the standard hierarchical registration scheme, which is recommended for general use in inter-subject structural brain registration within FNIRT, is used. This is given in Table 3.1. As FNIRT only provides a relatively coarse registration, with a control point spacing of 10mm by default, this is extrapolated to provide a more accurate 5mm control point spacing. Both a highly regularised, and a lower regularised 5mm FNIRT configuration are chosen for comparison. The hierarchical scheme consists of 4 levels of sub-sampling resolutions and control point spacings. At each of these levels, there are two separate sub-levels with differing amounts of regularisation and image pre-smoothing.

Image Smoothing

The images are smoothed prior to subsampling. This is both to improve the SNR, and provide a scale-space representation of the image data. Only image features of a specific scale are preserved by the smoothing [191]. The level of image smoothing alters the image data in $\mathbf{x}$ and $\mathbf{y}$ by removing noise and blurring features. This subsequently affects the inference on the registration model as the VD factor estimates the covariance between voxels in the residual, where smoother images result in a lower VD factor. Greater levels of image smoothing improves the accuracy of the 1st order Taylor series approximation of the transformation

function, $\mathbf{t}(\mathbf{x}, \mathbf{w})$. Therefore, the level of image smoothing at each level of the hierarchical registration scheme will have an effect on the inferred mapping.

Some further experimentation was carried out with regard to image smoothing, and in the previously published results [169], a lower level of image smoothing is used. Subsequent experiments have shown that using a scheme with less smoothing provides slow convergence. For this reason, and for ease of comparison, the results presented in Chapter 5 use the standard FNIRT smoothing scheme, the details of which are given in Table 3.1.

As different scales of image features are visible at different levels of the hierarchical registration scheme, and each level will have varying residual magnitudes and degrees of freedom, all of which greatly affect the model fit, they will commonly require substantially different values of ϕ and λ.

3.5 Discussion and Conclusions

3.5.1 Discussion

This chapter has proposed a framework for inferring the level of spatial regularisation in non-rigid registration as part of a probabilistic inference scheme. There are two key advantages of formulating this model within a principled probabilistic framework; firstly it produces estimates of the uncertainty of model parameters. Chapters 6, 7 and 8 explore uses of the estimated transformation uncertainty. A second advantage is that this framework is generic and extensible, and it is possible to incorporate more complex transformation models, spatial priors or noise models. The next chapter investigates the use of a more complex noise model.

The VB framework permits model comparison using the variational free energy $\mathcal{F}$, which could be usefully applied in non-rigid registration, for example, to compare different control point spacings or spatial priors. Unfortunately the framework presented here has two factors that prove restrictive for model comparison: firstly the VD factor α, used to compensate for the spatial covariance of the residual image defined in section 3.2.2, would need to be fixed between methods, as this would otherwise alter the data. Additionally, the Taylor series expansion will give different uncertainties at different local minima. Due to these difficulties, Bayesian model comparison is not considered in this thesis.

3.5.2 Conclusions

This chapter has described a probabilistic framework for the tractable inference of regularisation parameters in high-resolution non-rigid registration. This has been developed as a generic registration framework, capable of using a range of transformation and regularisation models.

The next chapter describes an extension to this model that allows spatially varying estimates of noise. Chapter 5 presents a quantitative evaluation of both registration methods.

Chapter 4

Probabilistic Non-Rigid Registration Using Local Noise Estimates

4.1 Introduction

This chapter introduces a more flexible extension of the registration model presented in the previous chapter. Specifically, the model is extended to have spatially varying estimates of model noise. This allows the data driven inference of a spatially varying trade off between data fidelity and regularisation in non-rigid registration. This work has been previously presented [167].

4.1.1 Motivation

Intensity based registration methods require a cost function to measure image similarity, which is used to drive the optimisation. In the approach presented in Chapter 3, the sum of squared differences between the transformed source image, and target image is used. Regardless of the choice of image similarity metric, a problem common to all of these methods is the presence of ill-matching anatomical structures between the two images. Here, ill-matching anatomical structures are defined as those regions where a smooth and accurate mapping between two subjects, using a given transformation model, is difficult to obtain.

In the case of inter-subject brain registration, there is potential for the appearance of ill-matching anatomical structures. For example, there may be regions of the cerebral cortex that have a complex geometric variability between subjects, and therefore may be difficult to accurately estimate the correspondence for. There may even be some cortical structures that do not have homologues in all subjects.

The significance of these regions to non-rigid inter-subject brain registration, lies in the trade-off between the level of data fidelity and regularisation. Where a global value of these parameters is used, and a higher weighting is selected for data fidelity, well matching brain structures should be accurately registered. However, there may also be multiple regions where an unreasonable mapping, which may not be smooth, is inferred in an effort to reduce the cost-function. While allowing such a mapping may provide a slight improvement in image similarity cost, it will not necessarily provide a more accurate mapping between subject anatomies. Alternatively, where a higher weighting is given to the spatial regularisation, globally smoother transformations will be inferred. This may under-estimate the mapping in some regions where an accurate matching would be possible given greater flexibility.

4.1.2 Previous Approaches

The use of a spatially varying trade-off between image fidelity and regularisation has been previously explored by several authors. Davatzikos [45] computed similarity gradients from a representation of the cortical surface of the brain, and a segmentation of the ventricles to drive the registration. A spatially varying set of elastic regularisation parameters are defined to allow for a range of deformation sizes in segmentation defined regions of the brain. Some incorporated explicit prior information, Lester et al. [114] define a spatially varying likelihood weighting function and regularisation form derived from the type of anatomical tissue at a location. They suggest that the weight assigned to the likelihood, should be varied according to prior assumptions about the relevance of the data in different regions of the image. Regularisation should be used to ensure smooth mappings in regions with a low data likelihood, which draw influence from more trusted regions. Recent approaches also use prior knowledge of tissue types from segmentations to define the level of regularisation [173]. Such approaches require a substantial amount of prior knowledge, and do not adapt to the presented data.

Data-driven spatially varying regularisation approaches have previously been approached in terms of anisotropic smoothing of image similarity gradients according to the homogeneity of image information, as initially proposed by [130]. This has since been adopted to medical image registration [83]. Such approaches still require a proportional trade-off between data fidelity and regularisation to be defined, as well as the smoothing parametrisation.

Other data driven approaches have focused on the concept of image feature saliency [101]. In such approaches, image regions can be identified that are more likely to be unique, and can therefore be weighted more heavily in the similarity function as they should provide a more useful description of image matching. Huang et al. [92] proposed such an approach to identify salient features in each image independently. However, as is the case with ill-matching structures, salient features in one image are not necessarily salient in the other. Luan et al. [117] identify the utility of each voxel according to a regional estimate of the joint image saliency, where the gradient of a mutual information similarity gradient in rigid registration is weighted higher if a pair of voxels is salient in both images. Tang et al. [178] use a local image reliability measure based on image structure and estimated noise levels, which was used to derive a spatially localised trade-off. A limitation of that approach is that only the information in the individual images are used to define the local regularisation weighting, which may still lead to problems dealing with ill-matching structures. The most thorough approach to be suggested is that by Ou et al. [132]. In their approach, image data is assessed and appropriately weighted based on the presence of mutually salient image features. This has been demonstrated to improve registration, although this approach is computationally highly expensive.

4.1.3 Proposed Solution

This chapter proposes the use of local measures of data fidelity as an extension of the previously proposed algorithm. This provides a mechanism for describing a spatially varying level of trust in the image data, and therefore the derived similarity-measure gradients in a region. Spatially varying estimates of noise should aid the inference of smoother warps in regions that are ill-matching, while allowing larger warps in regions where a better match is obtainable. Furthermore, this increases the validity of the Gaussian noise model, as the residual image is more likely to follow a locally Gaussian distribution as ill-matching structures are spatially localised. Incorporating such a procedure within the previously

described probabilistic registration algorithm is particularly beneficial, as a global level of spatial regularisation is inferred, and the estimated level of uncertainty in the mapping should be accordingly higher in ill-matching image regions. A similar extension is possible to the framework of Risholm et al. [146], which uses MCMC sampling rather than variational Bayes for inference. As described previously in section 3.1.2, approximate full Bayesian inference using VB is a preferential strategy to MCMC due to its computational efficiency.

4.2 A Generative Model of Image Registration with Local Noise Estimates

This method extends the probabilistic registration framework described in Chapter 3 by using a generative model for registration that describes spatially-varying noise estimates. As before, the generative model is written as:

$$\mathbf{y} = \mathbf{t}(\mathbf{x}, \mathbf{w}) + \mathbf{e} \tag{4.1}$$

where $\mathbf{y}$ is the target image, and $\mathbf{t}(\mathbf{x}, \mathbf{w})$ is the transformed source image, where the transformation is parametrised by $\mathbf{w}$. As before, although this approach is valid for any choice of transformation model, a B-spline FFD transformation model is used, where $\mathbf{w}$ describes the displacement of the B-spline control points.

The difference in this model, from the one presented in the previous chapter, lies in the definition of the noise, $\mathbf{e}$, which is assumed to be independently distributed Gaussian noise. However, whereas previously $\mathbf{e}$ describes a global level of noise, in this approach $\mathbf{e}$ is modelled as having a smoothly spatially varying precision:

$$\mathbf{e} = \mathcal{N}(0, \mathrm{diag}(\alpha \Phi^T \mathbf{b})^{-1}) \tag{4.2}$$

where diag takes a vector for use as the diagonal of a matrix, $\Phi = \{\phi_1, \phi_2, ..., \phi_L\}$, is a $L \times 1$ vector of locally calculated Gaussian noise precisions, where L is the number of noise components. α is the virtual decimation factor that compensates for spatial smoothness in the noise, as previously described in section 3.2.2. $\mathbf{b}$ is a $N_v \times L$ matrix representing the basis set that assigns the weighting of each ϕ across the N_v voxels in the image. $\mathbf{b}$ must represent a non-negative basis set that has a degree of spatial smoothness to allow estimates of the image gradients. In this work the basis set $\mathbf{b}$ is implemented using a set of equally spaced normalised

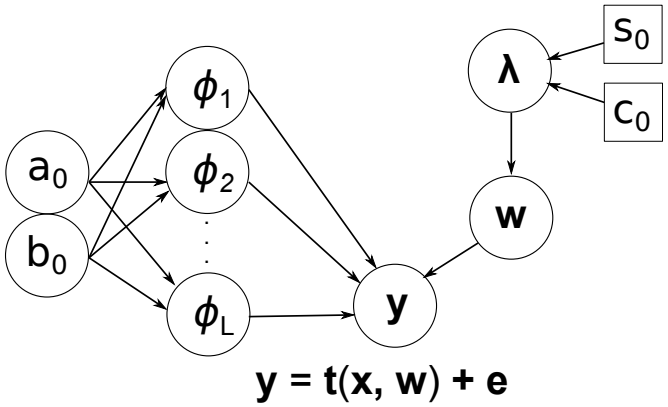

Figure 4.1: A graphical description showing the probabilistic dependencies of the registration model parameters. The variables in square boxes are constants, and the variables in circles are random variables. The Gaussian noise **e** of the model has multiple parameters $\phi_1, \phi_2, ..., \phi_l$.

Gaussian kernels. A graphical description of the extended registration model is given in Fig. 4.1.

Having a more complex model noise term leads to a different likelihood function. The log likelihood for this model is given by:

$$\log p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \Phi) = \sum_{l}^{L} \left(\frac{\alpha N_{v,l}}{2} \log \frac{\alpha \phi_l}{2\pi} - \frac{1}{2} \mathbf{m}_l^T \alpha \phi_l \mathbf{m}_l \right) \tag{4.3}$$

where $\mathbf{m}_l = (\mathbf{b}_{:,l})^{\frac{1}{2}} \circ (\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w}))$, here $\circ$ refers to the elementwise matrix product (Hadamard product). $\mathbf{m}$ is a $N_v \times 1$ vector. $N_{v,l}$ refers to the number of voxels weighted by ϕ. This likelihood function assumes an integer count in $N_{v,l}$ of all the voxels within a noise region. As the basis set may be smooth, in this work $N_{v,l}$ is approximated as the number of partial voxels weighted by ϕ_l, $N_{v,l} = \text{sum}(\mathbf{b}_{(:,l)})$.

4.2.1 Priors

The prior distributions of $\mathbf{w}$ and λ are the same as those in the Chapter 3 given in equations 3.5 and 3.6 respectively.

Independent prior distributions are placed on the hyper-parameters of each parameter in Φ, where for a given parameter l, ϕ_l is again described using a Gamma distribution:

$$P(\phi_l) = Ga(\phi_l; a_0, b_0) \tag{4.4}$$

where a_0 is the scale, and b_0 the shape parameter of the distribution. The prior distribution on noise precision is initially set to be uninformative with $a_0 = 10^{10}$, $b_0 = 10^{-10}$. However, once values of the ϕ parameters have been estimated, an informative prior can be derived. This prior can be calculated by fitting the current distribution of Φ with a Gamma distribution. The scale and shape parameters of the fitted Gamma distribution can then be used as a_0 and b_0.

4.3 Model Inference

As before, VB is used for inference on the parameters of the registration model and the mean-field approximation is required:

$$p(\mathbf{w}, \Phi, \lambda|\mathbf{y}) \sim q(\mathbf{w}, \Phi, \lambda) = q(\mathbf{w})q(\Phi)q(\lambda) \tag{4.5}$$

where $q(\Phi) = \prod_l^L q(\phi_l)$.

The negative variational free energy $\mathcal{F}$ for this model is defined as:

$$\mathcal{F} = L_{av} - D_{KL}(q(\mathbf{w})\|p(\mathbf{w})) - D_{KL}(q(\lambda)\|P(\lambda)) - \sum_l^L D_{KL}(q(\phi_l)\|P(\phi)) \tag{4.6}$$

where L_{av} is the expectation of the log-likelihood given in equation 4.3, with respect to the Gaussian noise model with spatially varying precision given by $\Phi^T\mathbf{b}$. $D_{KL}(q\|p)$ is the Kullback-Leibler (KL) divergence [108] between the approximate posterior, q, and the prior, p, for each of the groups of model parameters.

4.3.1 Inference on Noise Parameters

The posterior distribution for the noise precisions is approximated using independent Gamma distributions, where $q(\phi_l) = Ga(a_l, b_l)$. The updates for the hyper-parameters are given below:

$$b_l = b_0 + \frac{\alpha N_{v,l}}{2} \tag{4.7}$$

$$\frac{1}{a_l} = \frac{1}{a_0} + \frac{\alpha}{2}(\mathbf{r}_l^T\mathbf{r}_l + \mathrm{Tr}(\boldsymbol{\Upsilon}^{-1}\mathbf{J}_l^T\mathbf{J}_l)) \tag{4.8}$$

where $\mathbf{J}_l$ is the $N_c \times N_v$ Jacobian matrix of partial derivatives of the transformation parameters, calculated on the source image that has been weighted by the basis

component $\mathbf{b}_{(:,l)}$. $\mathbf{r}_l = ((\mathbf{b}_{(:,l)})^{\frac{1}{2}} \circ \mathbf{k})$, where $\mathbf{k}$ is a $N_v \times 1$ vector containing the difference image, $\mathbf{y} - \mathbf{t}(\mathbf{x}, \boldsymbol{\mu})$.

4.3.2 Inference on Transformation Parameters

The approximate posterior distribution on $\mathbf{w}$ is normally distributed, $q(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Upsilon}^{-1})$. The updates for these parameters are given as:

$$\boldsymbol{\Upsilon} = \sum_{l}^{L} \alpha \bar{\phi}_l \mathbf{J}_l^{\mathsf{T}} \mathbf{J}_l + \boldsymbol{\Lambda} \bar{\lambda} \tag{4.9}$$

$$\boldsymbol{\Upsilon} \boldsymbol{\mu}_{new} = \sum_{l}^{L} \alpha \bar{\phi}_l \mathbf{J}_l^{\mathsf{T}} (\mathbf{J}_l \boldsymbol{\mu}_{old} + \mathbf{r}_l) \tag{4.10}$$

where $\bar{\phi}_l$ is the expectation of the approximate noise distribution, $\bar{\phi}_l = \mathrm{E}[\phi_l] = a_l b_l$, and similarly $\bar{\lambda} = \mathrm{E}[\lambda] = sc$.

4.3.3 Inference on Regularisation Parameter

The approximate posterior distribution of λ is Gamma distributed, $q(\lambda) = Ga(\lambda; s, c)$. The hyper-parameter updates are as follows:

$$c = c_0 + \frac{N_c}{2} \tag{4.11}$$

$$\frac{1}{s} = \frac{1}{s_0} + \frac{1}{2} \left(\mathrm{Tr}(\boldsymbol{\Upsilon}^{-1} \boldsymbol{\Lambda}) + \boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{\mu} \right) \tag{4.12}$$

4.4 Implementation

This model is implemented as an extension to the previous probabilistic registration model in FNIRT [5], and the same approximations are made as in section 3.4.1. The hierarchical registration scheme is the same as before (given in table 3.1). The number of basis set components varies between levels of the hierarchical registration scheme, and from coarse to fine is given as: 1, 1, 8, 27, 27, 64, 64, 125. The basis set kernels have a full-width at half-maximum that is 50% of the size of the spacing between kernel centres. The size and form of the basis set was chosen empirically through experimentation to provide a visually reasonable level of localisation without undue computational burden.

The use of a local noise model for registration leads to an approximately five fold increase ($\sim$ 15hours), compared to using a global noise model ($\sim$ 3hours). However, as the majority of the computation is in calculating $\mathbf{J}_l^T\mathbf{J}_l$, which is separable, this could be made significantly faster if parallelised.

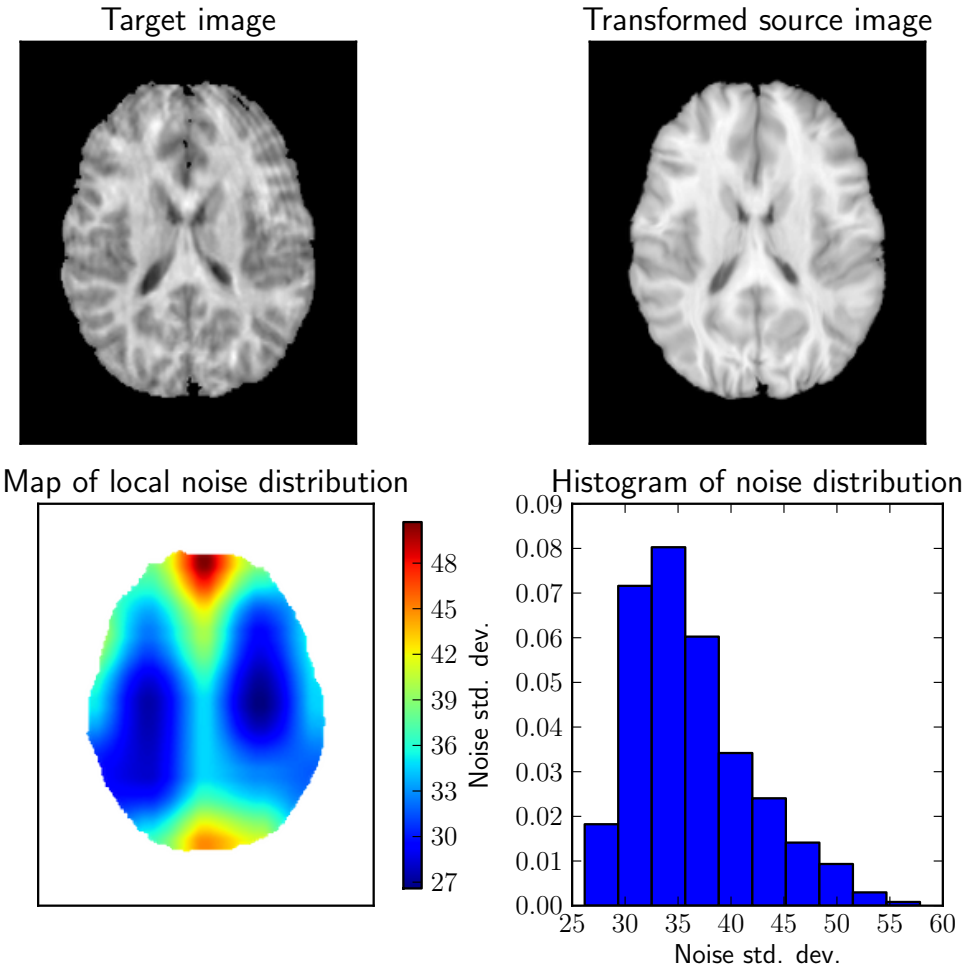

Figure 4.2: An example registration result. The top-left image shows the target image, the top-right, the transformed source image. The bottom-left shows the map of the local noise standard deviation for the example slice. The bottom-right is a histogram of the distribution of standard deviations of the noise across the whole brain. As can be seen, there is a reasonable degree of variation in the local noise distribution. For this slice, there is a much higher inferred level of noise and the top and bottom of the slice. This is because of the noise at the top of the image, and the difficulty in matching the sulci at the bottom of the image.

4.4.1 Registration Example

An example to illustrate the use of a local noise model is given in Figure 4.2. This shows one of the images in the IBSR dataset, described in section 5.3, registered to another and the resulting local noise distribution.

4.5 Conclusions

This chapter has proposed a generic probabilistic registration framework with a spatially varying noise model. The improved modelling of the residual error should provide a more appropriate balance between data fidelity and regularisation across the image. The limitation of this approach is the computational complexity required to estimate the L components of $\mathbf{J}^T\mathbf{J}$ weighted by the basis set. In the current implementation, this increases the computational burden by approximately a factor of 5 for the hierarchical registration scheme considered. However, as each of these components are calculated independently, this could be parallelised to run on an appropriate hardware system to allow rapid registration.

The next chapter presents a full evaluation of the registration frameworks presented in this, and the previous chapters.

Chapter 5

Registration Validation

5.1 Introduction

This chapter introduces principled approaches for the quantitative validation of non-rigid medical image registration in the human brain. Data from two public image datasets are selected to explore the effects of signal-to-noise ratio, anatomical variability, and image contrast on the inferred registration model parameters. Subsequently, a quantitative comparison between the methods presented in Chapter 3, Chapter 4 and the FMRIB non-linear image registration tool (FNIRT), is performed in terms of registration accuracy and transformation smoothness.

5.2 Validation of Non-Rigid Registration

As was described in Chapter 2, there are are a variety of approaches for analysing morphometric differences between subjects that rely on non-rigid registration. To be confident in the validity of the inference of subject differences, the registration procedure needs to infer an accurate mapping between biological structures. Problematically, structural MR images themselves only contain a noisy surrogate for the tissue type at a location. In the case of inter-subject brain registration, this is unlikely to provide sufficient information for a perfect correspondence. Moreover, this implies image correspondence does not necessarily guarantee anatomical correspondence [42][151]. As has been eloquently described in [151], image similarity, which is often the measure being minimised, is a very poor method of registration validation. Having a high degree of image similarity does not guar-

antee the biological accuracy of the inferred registration. Furthermore, the "true" correspondence between anatomies is always unknown. For this reason, "gold-standard" measures have been derived to provide a means of comparing whether one particular mapping is biologically better than any other.

5.2.1 Gold standards

Simulations

For certain applications, the derivation of suitable biomechanical models have been used to provide gold-standard deformations [32][161]. The biological accuracy of these simulations is essential for such models to provide useful data for validation. The uncertainty surrounding the validity of any biomechanical model makes the use of such an approach questionable for the evaluation of registration accuracy.

Assessment of Anatomical Structural Overlap

In the absence of suitably realistic deformation fields, the assessment of the overlap of anatomical structures between the transformed source image and the target image can provide a measure of registration accuracy. Overlap in this case measures whether the same anatomical structure is present in each voxel between the two images. These anatomical structures need to be segmented to a sufficient level of accuracy to be suitable. Furthermore, the segmented structures need to be more complex than simply tissue classes to provide a useful measure [151]. Klein et al. [105] use a variety of assessment criteria for the accuracy of structural overlap and they note that these methods "gave almost identical results when corrected for baseline discrepancies". All measures being approximately equal, the target overlap metric is chosen as it allows a comparison with some results in Klein et al. The target overlap quantifies the intersection of a labelled region in two images, normalised by the size of the region in the target image:

$$TO = \frac{|A \cap B|}{|B|} \tag{5.1}$$

where TO is the target overlap, A is a labelled structure in the transformed source image and B is the corresponding labelled structure in the target image. $\cap$ refers

to the intersection of the structure in two images, and || indicates the number of voxels in the labelled region.

5.2.2 Other Measures of Registration Quality

It is highly unlikely that measures of anatomical structural overlap could have a sufficiently high resolution to describe the voxel-to-voxel relationship between images. For this to be possible, each voxel would require a unique label. Instead, these approaches commonly measure the overlap of regions with a similar structure/function. This lack of specificity means that there may be many possible mappings that provide the same, or very similar, gold standard measurements.

If the accuracy of several registrations were the same, the most appropriate registration to choose would be that which best preserves the relative positions of structures in the source image after transformation. Such a transformation should be smooth and with minimal or no image folding. These criteria can be considered as secondary objectives of a registration algorithm.

5.3 Materials

Two publicly available datasets were used in the evaluation of this method. The first is available from the internet brain segmentation repository (IBSR), which is from the Centre for Morphometric Analysis[1]. This data has been bias-field corrected and includes segmentations of both cortical and subcortical brain structures. An example slice from one subject is given in Figure 5.1

The IBSR dataset contains 18 subjects that have a variety of ages, between 7 and 71, with an average age of 38.4. There are also 4 scans of "juveniles". 4 of the subjects are female, the remaining 14 are male. The data were acquired with a variety of scan resolutions, 8 have a resolution 0.94 x 0.94 x 1.5mm, 4 are 0.84x0.84x1.5mm and 6 have 1x1x1.5mm. Additionally, the image contrast, signal-to-noise ratio (SNR), and presence of data artefacts is observably different between scans. This qualitatively appears to be linked to the scan resolution, where images with higher in-plane resolution have less noise, but more artefacts.

In the study by Klein et al. [105] they state "All of the algorithms performed worst on the IBSR18 set, whose images were acquired from various sources and are of varying quality". The variability between images, particularly in terms of image

[1] www.cma.mgh.harvard.edu/ibsr/

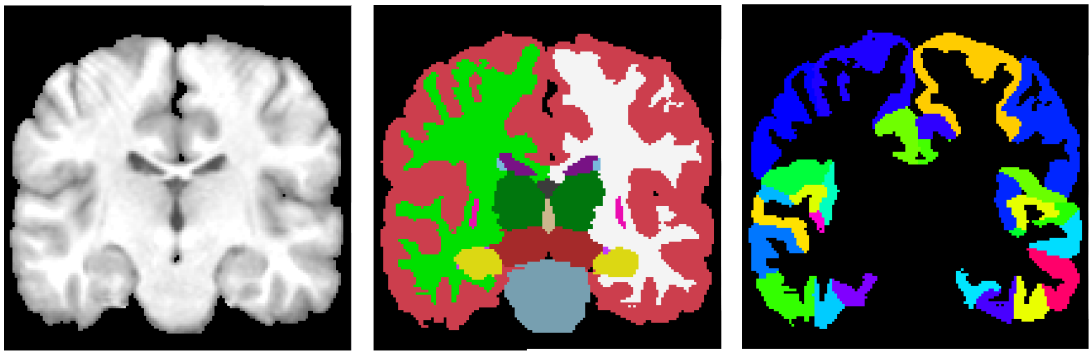

Figure 5.1: An example slice of a subject in the IBSR dataset. The left hand image shows the structural MR, the centre image shows the subcortical structure labels. The right image shows the segmentation map of the cortical regions in this slice.

contrast, makes this dataset an interesting one to work with, allowing a more thorough exploration of the difficulties in registration. This image variability and the provision of subcortical structure labels is why the IBSR database is used in almost all the experiments preferentially to the NIREP dataset, which was used in the previous evaluation found in [169].

The non-rigid image registration evaluation project (NIREP) [35] contains a set of 16 3D T1 weighted MR images of the human brain, taken from 16 healthy subjects. These images have been pre-processed by correcting for bias fields and were created by averaging 3 scans, and thus have unusually high SNR. As these images have uncommonly high SNR, and do not include subcortical structure labels, the NIREP data is only used in the signal-to-noise experiment as it allows the effects of higher SNR to be observed.

Prior to conducting the experiments, each image is affinely aligned to the MNI152 atlas using FLIRT [98] with 9 degrees of freedom and a correlation ratio cost function. These parameters were chosen to match the validation framework of Klein [105]. However, following some experimentation it was found that the best approach to registering brain extracted images with FLIRT is to use an image weighting function on the source and the target image. This weighting function is calculated by blurring the brain mask using a Gaussian kernel with a standard deviation of 2.5mm. The advantage of such an approach is that the majority of empty voxels do not contribute to the cost function, except for those surrounding the edge of the brain that produce a useful effect. This yields a more accurate and robust affine registration. Each of the individual non-rigid registrations is initialised with a further rigid alignment between the MNI 152 aligned images to

Name	Inferred λ	Local Noise Estimates	Informative Prior on λ
FNIRT/FNIRT2	No	No	Yes
FNIRT-VB	Yes	No	No
FNIRT-VB-IP	Yes	No	Yes
FNIRT-VB-LN	Yes	Yes	No

Table 5.1: Summary of the registration algorithms being compared.

resolve any differences in pose, this is estimated using FLIRT with 6 degrees of freedom.

5.4 Overview of Experiments

In these experiments, five algorithms are compared, and a summary of these variants is given in Table 5.1. The proposed probabilistic registration frameworks are evaluated in comparison to the standard FNIRT in FSL [5], as this formed the basis for the implementation. FNIRT and FNIRT2 are identical except for the level of regularisation at the final hierarchical registration levels, where FNIRT2 uses comparatively less regularisation than FNIRT. The level of regularisation for the FNIRT methods, and the other details of the hierarchical registration scheme are given in Table 3.1. For each method, a fifth-order non-linear intensity mapping is estimated between the two images, as is standard in FNIRT. Each image in the IBSR dataset is registered to every other using all the registration algorithms. From these registrations, the variability in the inferred λ and ϕ parameters across a range of signal-to-noise ratios, hierarchical registration levels, and between individuals is examined for the proposed methods. Subsequently, the quality of the inferred registrations is evaluated in terms of structural overlap measurements, transformation complexity and the level of image folding.

5.5 Variability in Inferred Regularisation

5.5.1 Variability in λ Across Signal-to-Noise Ratios

Theoretically, more regularisation may be required in situations where the image information is of lower quality. This is to avoid the registration becoming under-constrained. The variability in the inferred level of λ was examined over a range

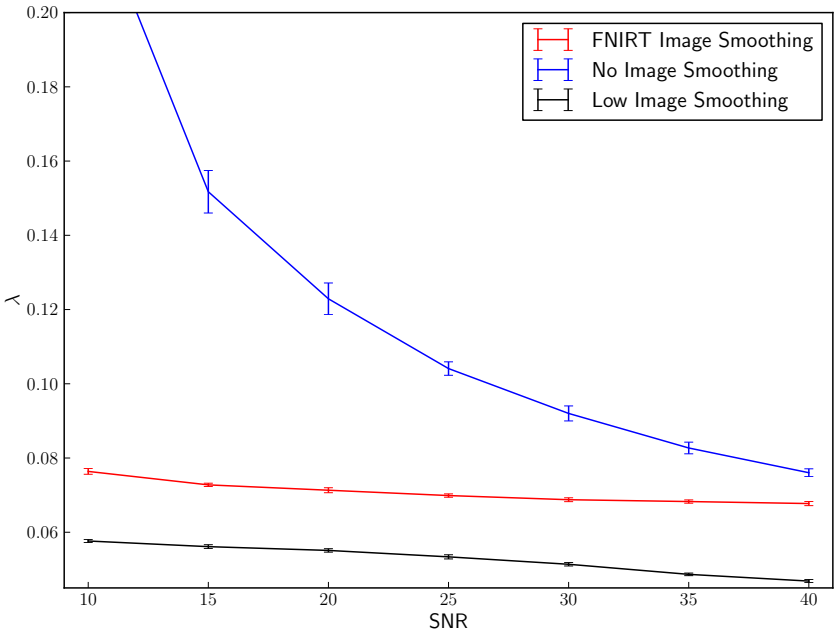

Figure 5.2: Variability of λ at the finest hierarchical registration level when registering images with a range of SNR. Two images from the NIREP database were registered using FNIRT-VB. Independent Gaussian noise was added to the source image at a range of signal-to-noise ratios to produce images between 10 (20dB) and 40 (32dB). The error bars show the standard deviation of λ across 10 instances of random noise. Three image pre-smoothing schemes were tested: the standard FNIRT scheme, a low pre-smoothing scheme (FWHM equal to the voxel size, and then half the voxel size), and a scheme with no image pre-smoothing. The plot shows that without smoothing the data, the algorithm is very sensitive to noise. There is a decrease in λ for higher SNR for all methods, although the change is quite minor in the case of pre-smoothing. This indicates that smoothing the data does a reasonable job at removing the effects of image noise. When using the low image smoothing scheme, the inferred λ values are substantially lower than with the other schemes.

of signal-to-noise ratios (SNRs). This is achieved by adding Gaussian noise to the source image, which had an original SNR of 40, to produce a set of images with a range of SNR between 10 (20dB) and 40 (32dB). The SNR of the target was estimated at 52 (34dB). Each noisy image is registered to the unmodified target image with the proposed framework. As the anatomy of both images remains the same, the difference in inferred regularisation between the registrations will only depend upon the SNR of the source image.

Noise is added to the source image as this mimics the common spatial normalisation procedure, where collected data is registered to some common average atlas. It should be noted that a more principled generative model would treat

the atlas as the source, as this should be part of the model, and deform the atlas to the observed images. However, such an approach requires the inversion of the inferred deformation field, which for certain transformation models, such as B-spline FFDs, may not be well defined.

Adding noise to the target image instead will lead to a slightly different behaviour. This is because gradients are calculated based on the source image, and noise in the source image is smoothed by the interpolation used to warp the source image.

10 noisy images were sampled for each SNR level to provide error bounds on the estimates. As image smoothing increases the SNR of the data, but preserves fewer small features, these effects are investigated here with three schemes: the original FNIRT scheme, a scheme employing no pre-smoothing and a lower image smoothing scheme. The inferred level of λ for each SNR level is shown in Figure 5.2. Where no image smoothing is used in the presence of noise, a very high λ is inferred. This is because the 1st order Taylor series approximation used by this model (described in section 3.3.3), which assumes the transformed image changes locally linearly with respect to the transformation parameters, is mostly invalid when the data is non-smooth. This results in the estimation of weak, and unreliable cost function gradients. For data with good SNR, the FNIRT image smoothing and no image smoothing produce similar results. Whereas, using little pre-smoothing leads to very low λ. For this reason, and its robustness to noise, the FNIRT image smoothing scheme is selected for all further experiments.

5.5.2 Variability in λ and ϕ Across Subjects and Hierarchical Registration Scheme Levels in FNIRT-VB

Examining the distribution of inferred λ and ϕ over a set of example registrations provides an illustration of the behaviour of the algorithm. Each level of the hierarchical registration scheme requires a different amount of regularisation depending on the control point spacing, the image pre-smoothing and the image sub-sampling. These factors affect either the transformation model, or the data to be registered. Additionally, each pair of individual brain images may require a different trade off between regularisation and data fidelity due to anatomical or image acquisition differences.

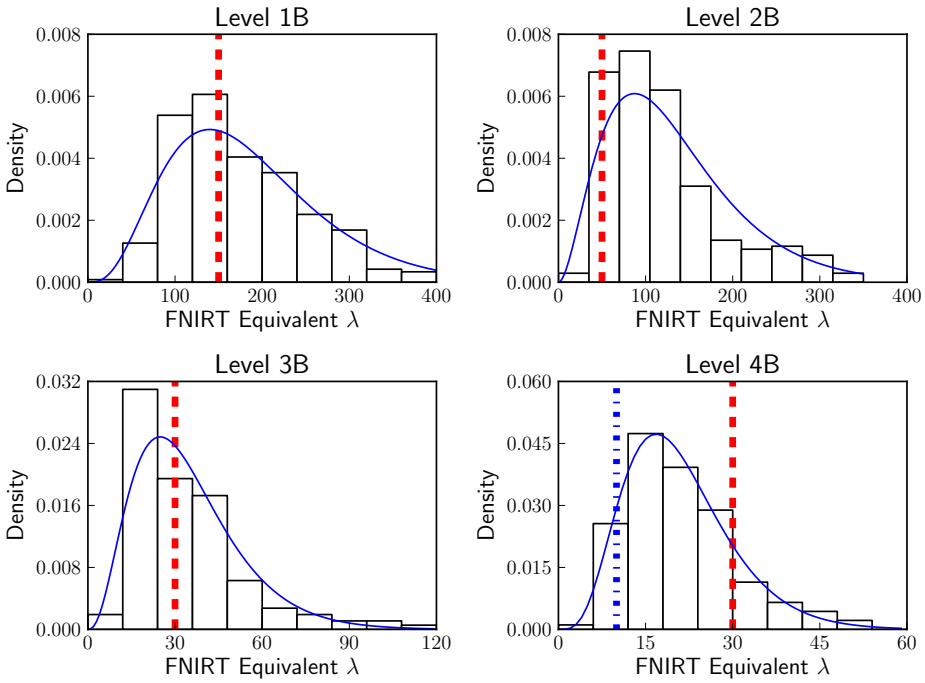

Figure 5.3: Histogram and fitted gamma distribution plot of the inferred FNIRT equivalent values of λ, over the 306 pairwise registration of the IBSR database using FNIRT-VB for levels 1B (coarsest), 2B, 3B and 4B (finest). The y-axis shows the normalised population density. The dashed red line indicates the λ value utilised in the original FNIRT configuration, the dash-dot blue line for level 4B indicates the low regularisation scheme for the finest levels. Note the similarity between the mode of the distributions to the hand-defined FNIRT values.

FNIRT equivalent λ distribution

The distribution of FNIRT equivalent λ (FEλ) values inferred by FNIRT-VB for each level of the hierarchical registration scheme, given in Table 3.1, is shown in Figure 5.3. As described in section 3.4.2, FNIRT encodes the trade-off of λ and ϕ into a single parameter that is referred to as FNIRT λ in this thesis. FEλ is calculated to give the value of the FNIRT λ parameter that would describe the same trade-off as the inferred λ and ϕ parameters. This allows comparison with the original FNIRT configuration. The inferred distribution of FEλ has a wide variability between subjects at each stage of the registration. As FEλ is the proportion of spatial regularisation to SSD, it also accounts for the noise in the model residual. Therefore the variation in FEλ is likely to be driven by either anatomical variability, or difficulties in estimating the intensity mapping.

Distribution of λ

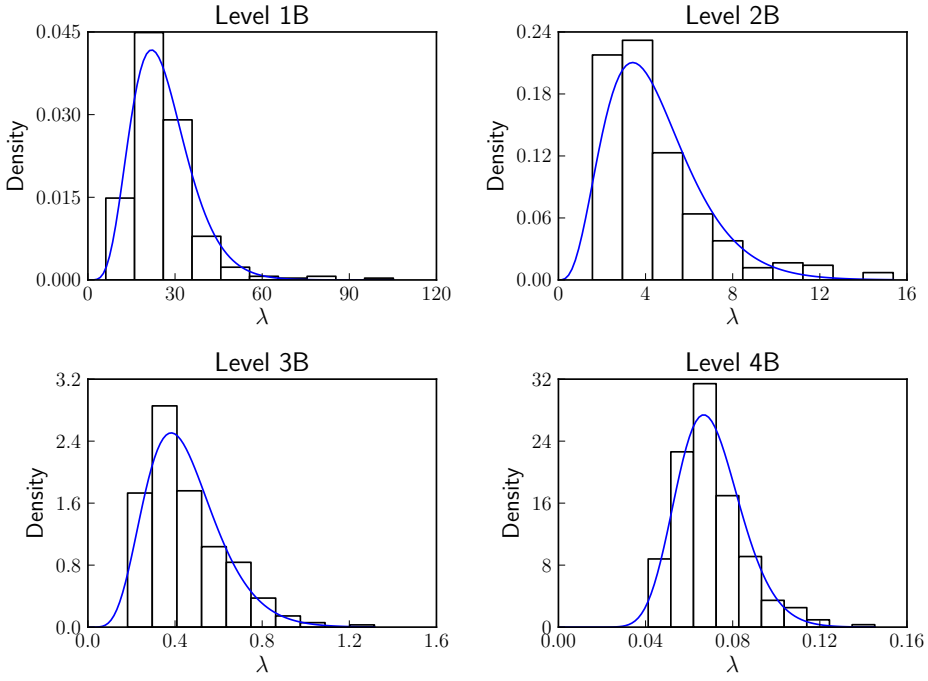

Figure 5.4: Histogram illustrating the distribution of inferred spatial precision (λ) across the 306 registration pairwise registration from the IBSR database using FNIRT-VB for levels 1B (coarsest), 2B, 3B and 4B (finest). The y-axis shows the normalised population density. The blue line shows the fitted Gamma distribution. The distribution of λ strongly resembles the estimated Gamma distribution, and shows no significant differences according to Kolmogorov-Smirnov test at the 5% significance level.

The distribution of λ values inferred across the 306 registrations in the IBSR database with FNIRT-VB is given in Figure 5.4. The λ value indicates the strength of the spatial prior. The distribution of λ is broad for each level, varying by a factor of 2 from the mean for most of the levels of the hierarchical registration scheme. However, the population is still well modelled as a Gamma distribution, indicating a single λ population. As λ shows little variability due to the SNR of the data after smoothing, it can be surmised that the majority of this variation is due to anatomical variability.

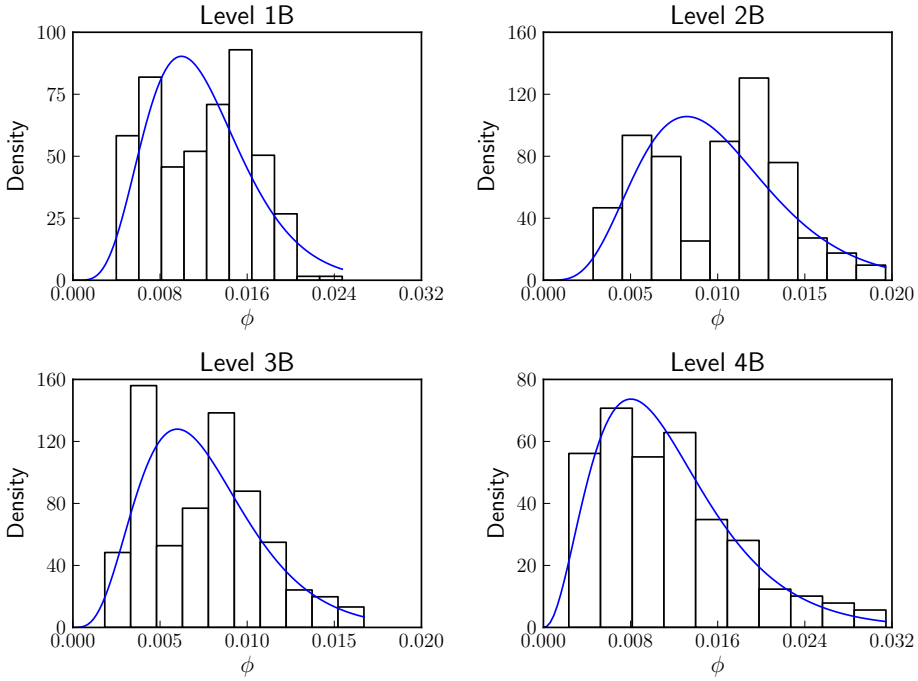

Figure 5.5: Plot illustrating the inferred noise precision (ϕ) across the 306 registration pairwise registration from the IBSR database using FNIRT-VB for levels 1B (coarsest), 2B, 3B and 4B (finest). The y-axis shows the normalised population density. The blue line shows the fitted Gamma distribution. The distribution of ϕ at each level of the hierarchical registration scheme is significantly different from the fitted gamma distribution (Kolmogorov-Smirnov test at the 5% significance level) and appears to be multi-modal.

Distribution of ϕ

The distribution of inferred noise precisions (ϕ) across the 306 registrations in the IBSR database using FNIRT-VB is given in Figure 5.5. The distribution appears to be multi-modal, which implies the distribution is a mixture of ϕ populations.

The multi-modal nature of the distribution of ϕ can be further investigated. Figure 5.6 shows ϕ plotted against λ in separate colours denoting the "resolution group" of the source image. The resolution group refers to the in-plane resolution of the original image before re-sampling. As mentioned in section 5.3 these different acquisitions have varying contrast, SNR and artefacts. In each group there is a trend for lower λ with higher ϕ. However, for all of the images where the source image was from the lowest resolution group the values of ϕ are much lower, although the λ values are similar.

The reason for the lower ϕ for this group is most likely related to the complex

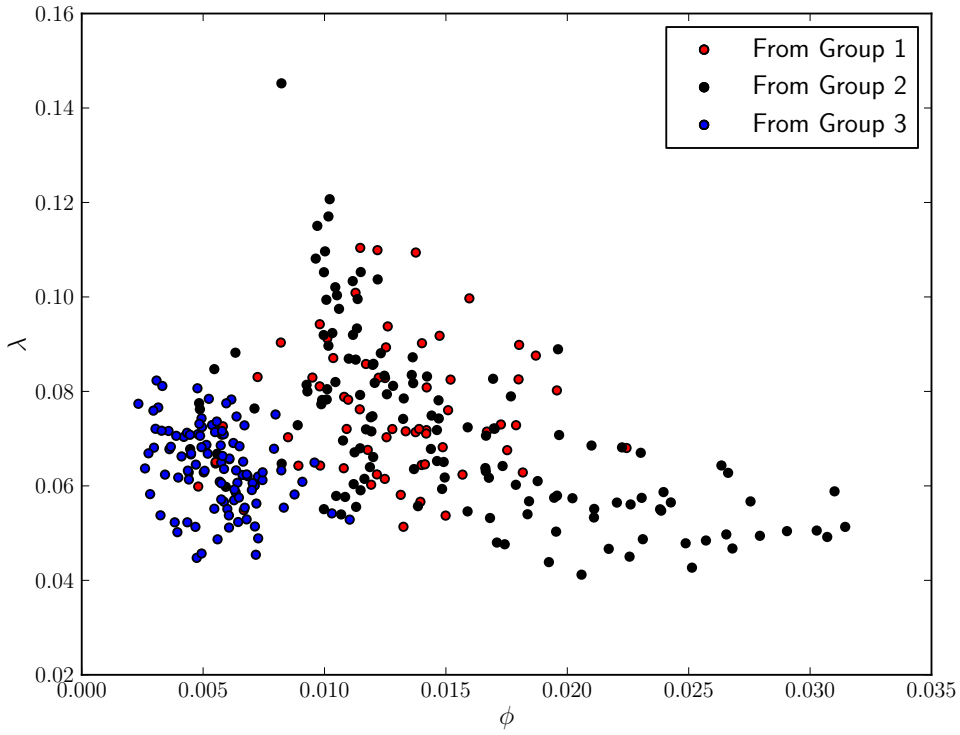

Figure 5.6: Scatter plot illustrating the relationship between ϕ and λ for the IBSR dataset at the finest level of the hierarchical registration scheme. The "resolution group" of the source images are shown by plots in different colours, with 1 being the highest resolution (0.84x0.84x1.5), 2 is slightly lower (0.94x0.94x1.5) and 3 is (1x1x1.5.). There is a clear relationship between ϕ and λ for all groups. Where the source image is in group 3, the inferred value of ϕ is much lower.

intensity mapping function required to map another image to the same contrast as the source. It is unlikely to be due to image noise that was not resolved by smoothing as similar λ values for all registrations are inferred, which implies that the image gradients are sufficiently visible in the source image.

5.5.3 Informative Prior Distribution

The ideal situation when multiple subjects are being co-registered, or spatially normalised, is that an informative prior distribution for λ, and/or ϕ can be inferred based on the current distribution of estimated parameters. Such a situation is currently computationally infeasible as multiple registrations would need to be carried out simultaneously, and iterate in step with each other. The potential benefit of an informative prior is that the inference of parameter values can be

regularised, giving a better conditioned inference scheme, and thus reducing the probability of outliers

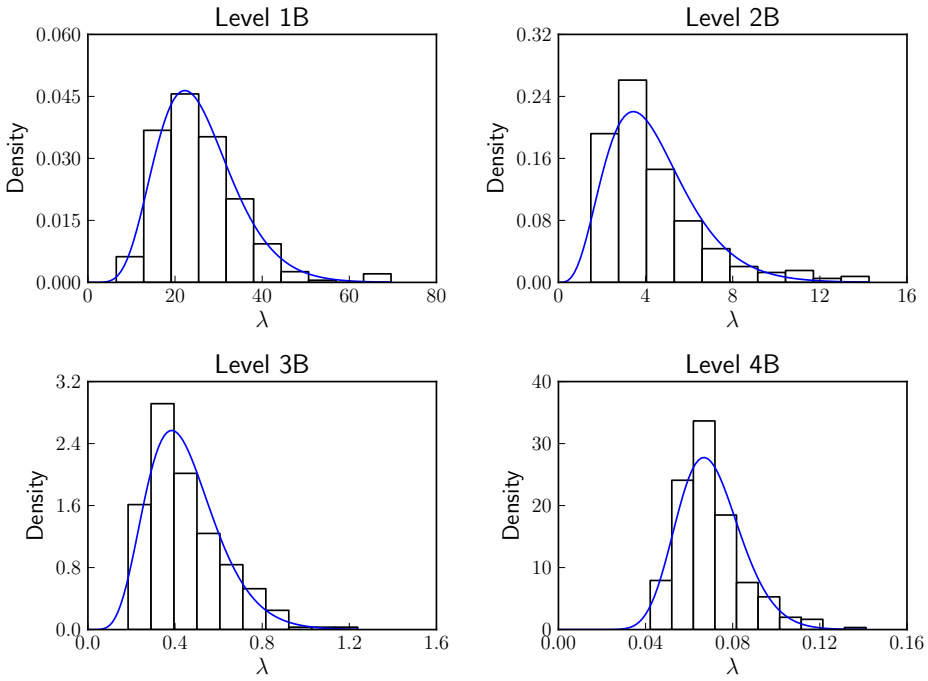

Figure 5.7: Histogram illustrating the distribution of inferred spatial precisions (λ) across the 306 registration pairwise registration from the IBSR database using FNIRT-VB-IP for levels 1B (coarsest), 2B, 3B and 4B (finest). The y-axis shows the normalised population density. Using an informative prior on λ removes some of the outliers that are present in the coarser levels of the hierarchical registration scheme, but makes very little difference to the λ at the finest level. The effects of the informative prior do not produce a statistically significant effect on the inferred λ distribution (Wilcoxon rank-sum at 0.05 level)

.

Informative priors can be estimated based on the distribution of converged estimates of λ, and ϕ. Figures 5.4 and 5.5 illustrate the distribution of inferred λ and ϕ parameters across the 306 registrations in the IBSR database using FNIRT-VB. The fitted Gamma distributions, where this provides an appropriate model of the population, can be used as an informative prior in the registration model. As shown in Figure 5.5, the distribution of ϕ is not well modelled by a Gamma distribution. This provides a caution against the use of an informative prior for ϕ, as it may vary significantly across acquisition types, whereas λ is likely to be more stable.

In these experiments, all 306 registrations were re-run using an informative prior for λ (FNIRT-VB-IP), based on the original 306 registration conducted with an uninformative prior. The distribution of λ values given by FNIRT-VB-IP is shown in Figure 5.7. The effects of using an informative prior in this manner appears to be minimal from the inferred λ values. However, the effects are likely to increase if the population of λ was modelled at each iteration, rather than constructing a prior from a converged set of λ values. Nevertheless, within the current approach, there is no evidence that λ requires an informative prior, and therefore its value is sufficiently well supported by the data. The FNIRT-VB-IP results are not presented in the quantitative validation as they are almost identical to the results of FNIRT-VB.

5.5.4 Variability in λ and ϕ Across Subjects and Levels of the Hierarchical Registration Scheme in FNIRT-VB-LN

Distribution of λ

Figure 5.8 shows the inferred λ distribution from FNIRT-VB-LN. The inferred λ values are slightly higher than those inferred using FNIRT-VB, however the distributions are not significantly different (Wilcoxon rank-sum at 0.05 level).

Spatial Distribution of ϕ

Due to the large number of ϕ parameters, and their spatial association, it is more appropriate to view the distribution of ϕ parameters as a map. The mean and standard deviation of ϕ across the IBSR registrations is given in Figure 5.9. The local noise model estimates higher ϕ values in the homogeneous white matter regions, and lower ϕ in the cortex. The inferred residuals in the white matter will tend to have relatively low magnitude ($1/\phi$), as the transformation parameters are marginalised over their posterior distributions in a homogeneous region. Whereas cortical regions are less homogeneous, and geometrically complex, making them harder to register and leading to larger residuals. The large standard deviation in ϕ across subjects in the white matter, compared to the cortical regions, is most likely related to how well the intensity mapping between images is estimated. This is because for any mapping in the posterior distribution of transformations, if the intensity mapping is poor, the residual is always large in homogeneous

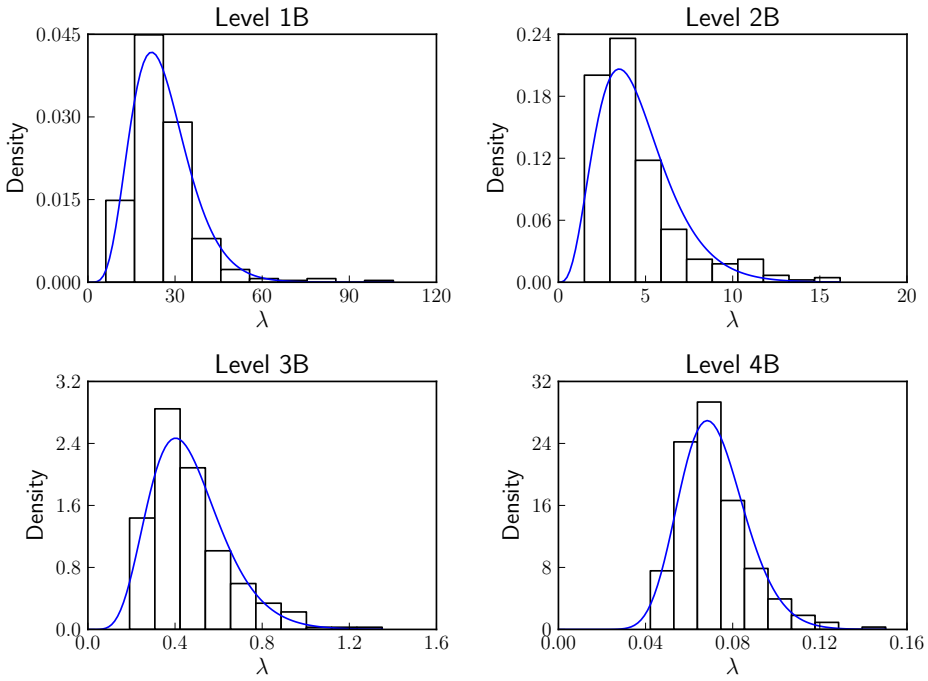

Figure 5.8: Histogram illustrating the distribution of inferred spatial precisions (λ) across the 306 registration pairwise registration from the IBSR database using FNIRT-VB-LN for levels 1B (coarsest), 2B, 3B and 4B (finest). The y-axis shows the normalised population density. The blue line shows the fitted Gamma distribution. The distribution of λ strongly resembles the estimated Gamma distribution, and shows no significant differences according to Kolmogorov-Smirnov test at the 5% significance level.

regions. Whereas in heterogeneous regions, the size of the residuals has greater dependence on the shape of the posterior transformation distribution, which in this framework draws image information from the source image only and therefore is less affected by the intensity mapping.

5.6 Validation of Inferred Registration Mappings

5.6.1 Registration Accuracy on Subcortical Labels

Figure 5.10 illustrates the overlap in subcortical segmentations using the different approaches. As can be seen, FNIRT-VB and FNIRT-LN2 achieve a similar level of overlap on subcortical structures to the original methods. However, in regions with more geometrically complex differences, such as the cerebral cortex and

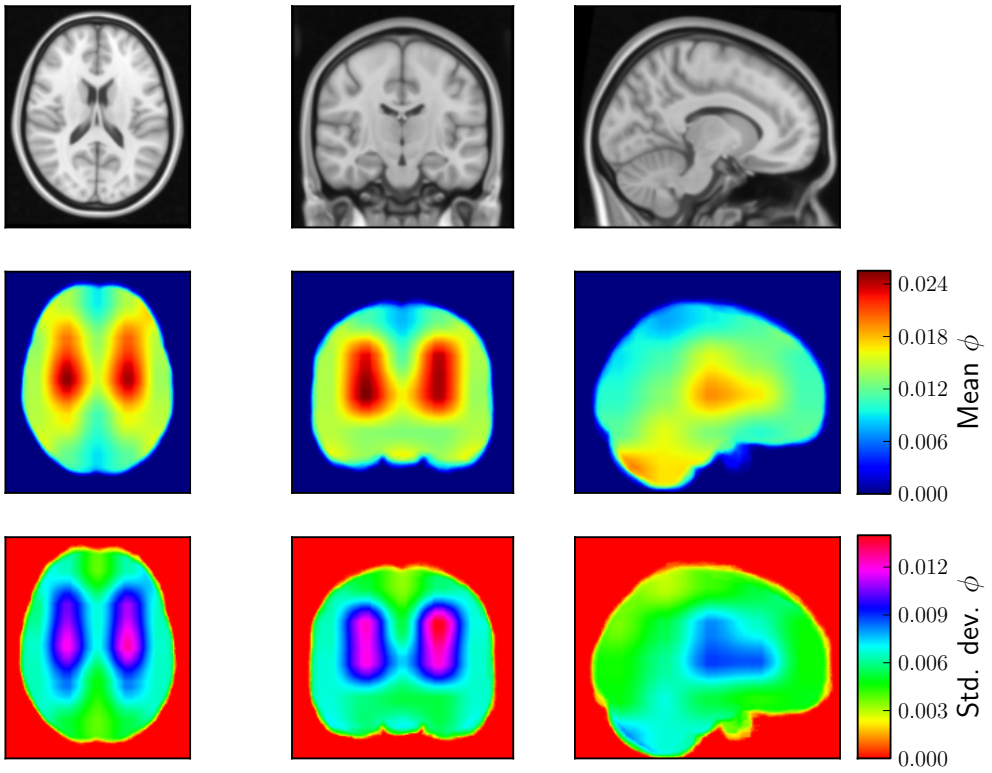

Figure 5.9: Maps of the mean and standard deviation of ϕ (inverse variance of the model noise residuals) across the 306 registration pairwise registration from the IBSR database using FNIRT-VB-LN. The results from the final level of the hierarchical registration scheme are shown. The top row shows the corresponding slice of the MNI152 atlas that the IBSR data is affinely registered to. The highest mean ϕ is understandably in the white matter regions, with decreasing ϕ towards the cortex. The regions with the most variation are the white matter and the cerebellum.

white matter, it can be seen that FNIRT2 achieves a higher level of overlap. Conversely, on other less complex structures such as the hippocampus and the thalamus, FNIRT-VB-LN performs best, followed by FNIRT-VB. Interestingly, FNIRT2 performs worst on some structures such as the pallidum, caudate and amygdala. This illustrates that under-constraining the registration may lead to poor registration of certain brain regions.

The differences between FNIRT-VB and FNIRT2 lies in the trade-off between λ and ϕ. As shown in Figure 5.3, FNIRT-VB mainly differs by having more regularisation at the second coarsest, and the final resolution level, although this varies depending on the data. However, on average this extra regularisation appears to provide a better registration of certain structures.

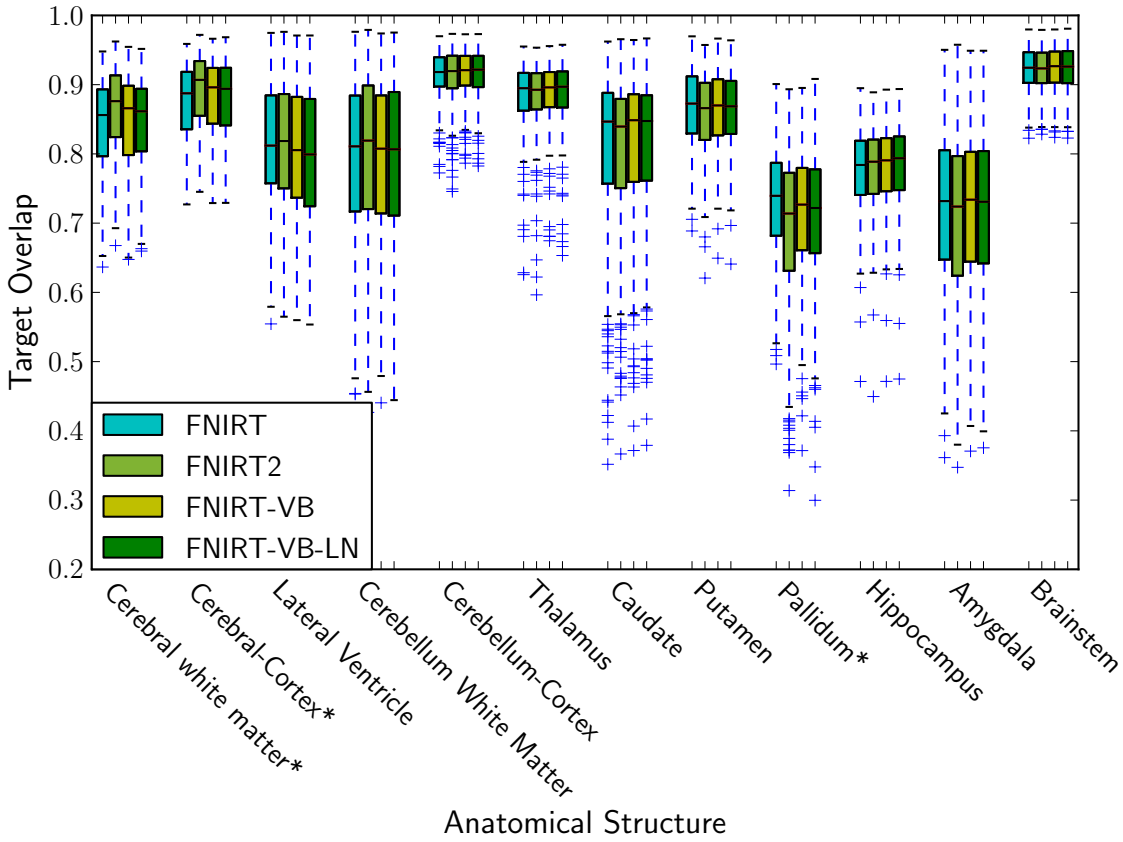

Figure 5.10: A boxplot showing the target overlap scores for a selection of subcortical labelled regions across the 306 registration pairwise registration from the IBSR database. The left and right label target overlap scores have been averaged for each subject. The 4 registration variants are displayed in different colours. The * at the end of a region name indicates a statistically significant difference in target overlap between the 4 populations, as measured by one-way analysis of variations (ANOVA), $p < 0.05$). The majority of structures show similar levels of overlap for all methods. FNIRT2 (lower regularisation) achieves significantly better overlap for the cerebral white matter and the cerebral cortex than the other methods, although it does significantly worse in the Pallidum.

There are few differences between FNIRT-VB and FNIRT-VB-LN in terms of overlap, none of which are statistically significant (Wilcoxon rank-sum 5% significance level). FNIRT-VB-LN is slightly better at registering the hippocampus and worse at the lateral ventricle. The inferred values of λ are similar for FNIRT-VB and FNIRT-VB-LN, so the majority of any differences will be because of the spatially varying ϕ. Although FNIRT-VB-LN infers a lower value of ϕ is in the cerebral cortex, a very similar level of overlap is recorded. Conversely, a high ϕ is found around the hippocampi and the lateral ventricles, the former of these is

well registered and the latter is less so.

5.6.2 Registration Accuracy on Cortical Labels

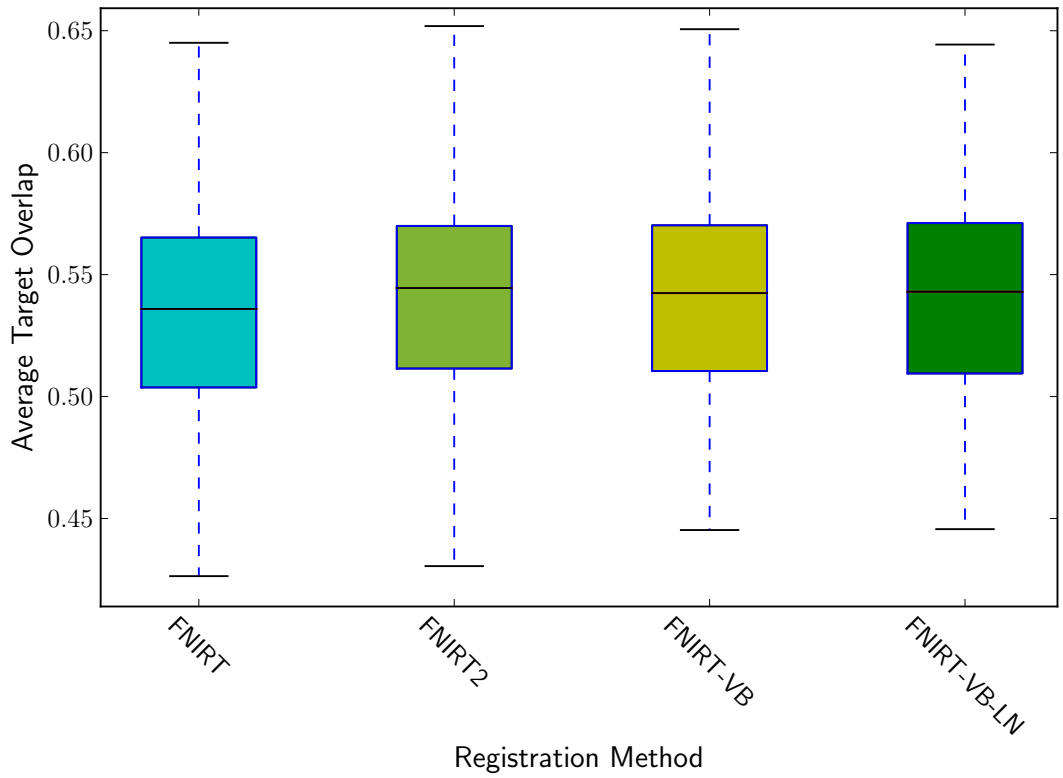

Figure 5.11: A boxplot showing the average target overlap for the cortical labels in the IBSR dataset across the 306 registrations. No statistically significant differences between methods were detected in the distribution of average target overlaps (one-way ANOVA, $p < 0.05$). The overlaps are quite consistent across all methods. Both the FNIRT-VB approaches obtain a better minimum overlap than the original methods, although FNIRT 2 shows a marginally higher median overlap. The overlap of cortical structures is expected to be lower than that of the subcortical structures because of the geometric complexities in the cortical folds.

The IBSR dataset also contains 96 cortical labels, the sizes of which are of approximately the same order. Therefore, it is reasonable to present results as an average of the target overlap for each subject, which is presented in Figure 5.11. Here, FNIRT2 achieves the highest average overlap (0.5445), although the minimum overlap is lower than the FNIRT-VB approaches. The minimum overlaps given by FNIRT and FNIRT2 occur for the data that the FNIRT-VB

approaches provide a more regularisation heavy trade-off. This additional regularisation appears to avoid the over-fitting of the registration algorithm, which may cause mis-registration in some regions for certain images. An example of this is given in section 5.7. FNIRT-VB-LN gives a slight improvement in median overlap (0.543) over FNIRT-VB (0.5425) but has a larger inter-quartile range. FNIRT with higher regularisation achieved the lowest overlap (median 0.536).

These cortical label results would be directly comparable with the results presented in Klein et al. [105] if the same pre-processing was used. However, as a more rigorous affine alignment is used, the overlaps show better results for all methods than any of the algorithms in that study. This is true to the extent that the lowest overlaps in these experiments are still better than the medians of some methods in that study. This highlights the benefits of doing an accurate global registration.

5.6.3 Registration Smoothness

Interpretation of Violin Plots

Violin plots provide a more detailed illustration of a set of data distributions, compared to the more standard boxplot. They feature a superposition of a kernel density estimate on a boxplot. The estimated density of the distribution for a particular value is given by the width of the "violin". This allows the visualisation, and comparison of the modes of a distribution as well as the medians and inter-quartile ranges given by the boxplot.

Bending Energy

Figure 5.12 shows the distribution of bending energy across all 306 registration for each method. As can be seen, for fixed levels of regularisation the bending energy forms a tight distribution. This is because a fixed ratio of bending energy to image similarity is expected for all registrations. When the level of regularisation is inferred, a much larger spread of transformation complexity is obtained. FNIRT infers the least complex transformations, but achieves the lowest level of overlap. FNIRT-VB and FNIRT-VB-LN infer transformations with a wide range of bending energies. Their modes and medians are quite similar, and substantially lower than those of FNIRT2.

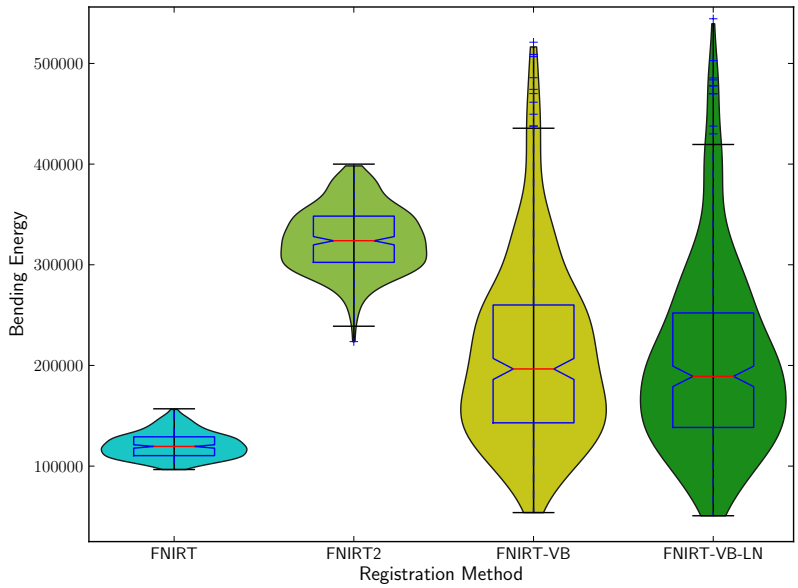

Figure 5.12: Violin plots of the level of bending energy across the different non-rigid registration methods for the 306 registration of the IBSR dataset. Bending energy provides a measure of transformation complexity across the different non-rigid methods. The fixed regularisation FNIRT approaches have a very tight distribution, implying a similar level of complexity used to map between all pairs of individuals. Conversely, where λ is inferred from the data, the level of bending energy takes a much wider range of values, as the regularisation trade-off is data dependent.

Negative Jacobians

The use of a small-deformation transformation model means that diffeomorphic registrations cannot be guaranteed. FNIRT has the facility to refit the deformation field to remove any negative Jacobians, but in this evaluation this is not performed. This allows the methods to be more fully compared in terms of the deformation field smoothness imparted by the regularisation.

The percentage of folded voxels in the transformation can be used as a further measure of transformation complexity. Figure 5.13 shows the percentage of folded image voxels across the set of IBSR registrations. The results are similar to the comparison of bending energies. FNIRT2 has the highest amount of folding, and FNIRT the lowest. FNIRT-VB and FNIRT-VB-LN have similar distributions of voxel folding, with a wide range of folding. The mode and median of these distributions is between that of FNIRT and FNIRT2. There are also some registrations using the VB methods where no folding occurs.

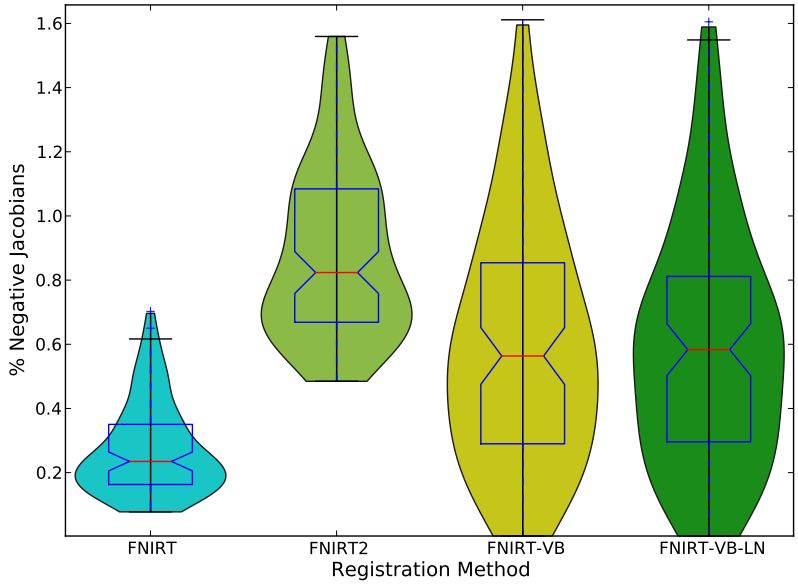

Figure 5.13: Violin plots of the percentage of negative Jacobians across the different non-rigid registration methods for the 306 registration of the IBSR dataset. FNIRT is capable of removing regions of folding by refitting the deformation field, but this is not performed here. The level of folding is almost always highest in FNIRT2. FNIRT achieves the lowest level of image folding, along with the lowest overlap. FNIRT-VB and FNIRT-VB-LN infer transformations with a wide range of folding. The mode and medians of which are lower than FNIRT2.

5.7 Example Registrations

5.7.1 FNIRT2 vs. FNIRT-VB

An example registration from the IBSR dataset illustrating the benefits of inferred regularisation is given in Figure 5.14. This is the image that achieved the lowest cortical label accuracy for FNIRT2. The use of a higher level of regularisation to data fidelity is demonstrated to provide a more accurate registration.

5.7.2 FNIRT-VB vs. FNIRT-VB-LN

An example registration from the IBSR dataset illustrating the benefits of a local noise model is given in Figure 5.15. This illustrative example was selected from the set of images where FNIRT-VB-LN performed better than FNIRT-VB. A similar level of λ is inferred for both registrations, yet FNIRT-VB-LN yields a more accurate registration as assessed by cortical label overlap.

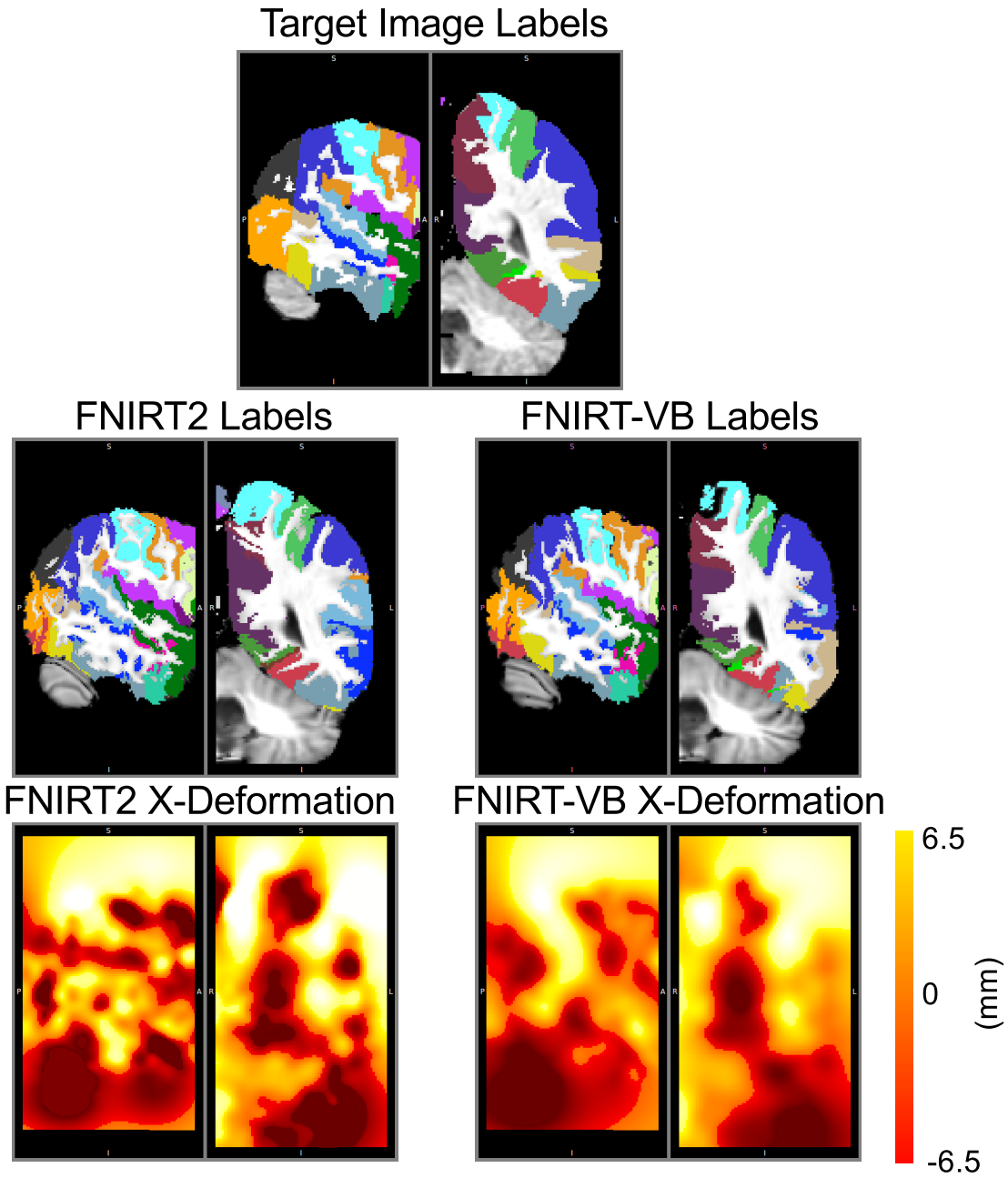

Figure 5.14: An example registration result comparing FNIRT2 and FNIRT-VB. The registration is from subject 10 to subject 13 of the IBSR dataset. Both subjects are in resolution group 3, which has been shown to infer lower values of ϕ. The top image shows a region of the target image with the different labels overlaid in colour. The second row shows the resulting transformed images with overlaid labels for FNIRT2 and FNIRT-VB. The third row shows the deformation field in the X-direction for these slices for both methods. The X-direction was chosen as it illustrated the greatest difference. FNIRT-VB achieves a better registration of the cortical labels, with an average of 0.453 as opposed to 0.430. FNIRT-VB infers a significantly higher FEλ for all levels of the hierarchical registration scheme, with a four-fold increase of regularisation on average. This leads to a much smoother deformation field, with much lower bending energy, 66,877 as opposed to 346,787.

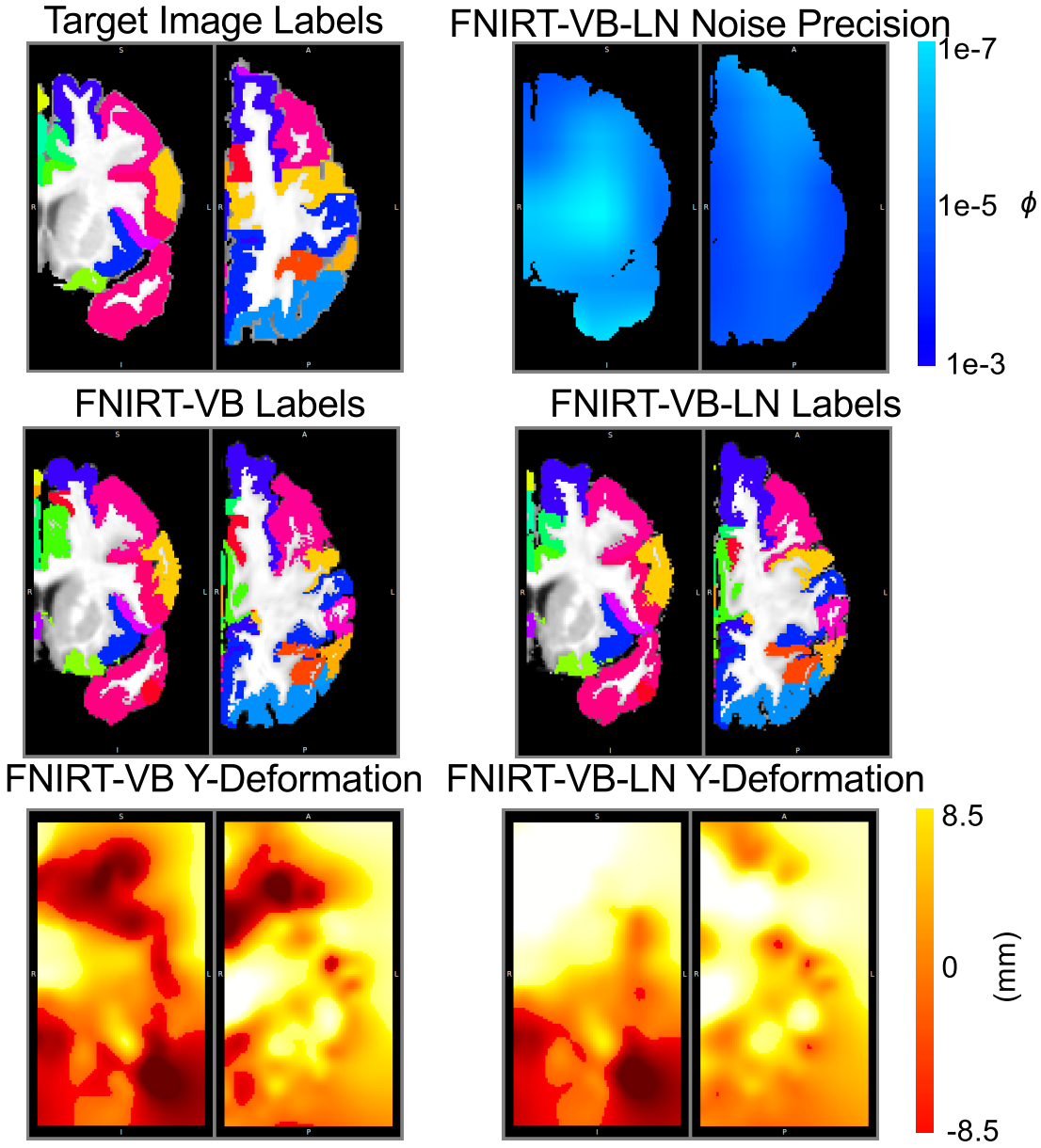

Figure 5.15: An example registration result comparing FNIRT-VB and FNIRT-VB-LN. The registration is from subject 12 to subject 2 of the IBSR dataset. Subject 12 is in resolution group 3 that has been shown to infer lower values of ϕ. The top left image shows a region of the target image with the different labels overlaid in colour. The top right image shows the inferred noise precision for FNIRT-VB-LN. The second row shows the resulting transformed images with overlaid labels for FNIRT-VB and FNIRT-VB-LN. The third row shows the deformation field in the Y-direction for these slices for both methods. The Y-direction was chosen as it illustrated the greatest difference. FNIRT-VB-LN achieves a better registration of the cortical labels, with an average of 0.483 as opposed to 0.466. FNIRT-VB-LN infers a slightly higher λ for each level, on average 14.3% higher. The local noise precision shows a range of ϕ values, which are low in most cortical areas. This leads to a smoother deformation of the cortex, with lower bending energy, 161,597 as opposed to 269,814.

5.8 Discussion

The registration experiments presented in this chapter have illustrated that the inference of the level of regularisation and data fidelity is important for inferring an accurate and smooth mapping between brain MR images. Although for some image pairs, lower levels of regularisation result in more accurate registrations, as measured by segmentation label overlap, this is not always the case. The most appropriate level of regularisation is data dependent, as is illustrated by the example registration in Figure 5.14. In this example higher regularisation provided a more accurate registration. The optimal trade-off of data fidelity and regularisation may also need to vary spatially as illustrated in Figure 5.10 and Figure 5.15. It must be noted that the Bayesian approaches to inferred regularisation tested here are not aiming to find the optimal trade-off based on the criteria of segmentation accuracy as has been previously proposed by cross-validation [202] or group-wise registration [183].

The FNIRT-VB framework has been demonstrated to yield a registration that is accurate for a range of anatomical structures across 306 separate registrations. It was shown to provide registrations with a variety of complexity as measured by bending energy and image folding. This means the registration algorithm is able to avoid the "one-size fits all" approach to registration complexity. Such an approach can lead to the inference of unjustifiably complex mappings, i.e. adding complexity for minimal benefit in the model fit, which can result in less accurate registrations. The variability in the inferred λ and ϕ values are mainly dependent on the image contrast and anatomical variability, with a minimal effect from image SNR under image pre-smoothing as indicated by Figure 5.2. The FNIRT equivalent λ values inferred from FNIRT-VB have similar modes to the default fixed values. An informative prior for λ was experimented with, and found to be unnecessary.

FNIRT-VB-LN has been shown to generally produce results quite similar to FNIRT-VB. The affects of a local noise model may be overly smoothed out by the use of the cubic B-spline model. Such a scheme may be more effective with a transformation model with greater degrees of freedom, or one that is less smooth.

One of the questions that still remains unanswered is how to pre-smooth the images. For this work, the default FNIRT level of image pre-smoothing is used. This provides robust and smooth registration across a range of SNRs. However, the optimal image smoothing is likely to depend on the specific application in question.

For all of the registration methods on this dataset, large deformations are inferred. This leads to image folding when using a small-deformation model such as the FFD transformation model used here. This framework could be implemented with a variety of transformation models, including those that are guaranteed to produce diffeomorphic mappings.

Most of the registrations in this chapter were performed on the IBSR dataset. This dataset features a range of ages, and image acquisition types. The differences in image acquisition leads to some complex intensity mappings between subjects, which may be difficult to model. The use of an inferred trade off, of regularisation and data fidelity, appears to compensate for this better than a fixed trade-off. This is well illustrated in Figure 5.6 where for images with complex intensity mappings the ϕ values are much lower than for the other data, but the λ remain similar. A further difficulty with this dataset is the labelled segmentation regions often take quite a blocky form. This is because of the manual labelling protocol by which they were derived. This segmentation "noise" will make high label overlaps difficult to achieve.

5.9 Conclusions

This chapter has introduced objective methods for measuring the accuracy of medical image registration. These were applied for the comparison of the algorithms presented in the previous two chapters, and FNIRT, on which the implementation of those methods is based. The IBSR dataset of 18 subjects with a range of acquisition types and subject ages was used for validation. This resulted in 306 registrations for each method. The distribution of inferred spatial and noise precisions, λ and ϕ respectively, were explored for the proposed registration models. A wide variability was demonstrated for both. ϕ was shown to be strongly affected by the acquisition type. The variation in λ is probably due to anatomical variability. The mode of the FNIRT equivalent λ parameter was shown to be similar to the default FNIRT schemes as shown in Figure 5.3.

The variational Bayesian registration methods on average are shown to produce similar, but slightly less accurate registrations to a fixed regularisation FNIRT scheme (FNIRT2) as measured by the overlap of subcortical (Figure 5.10), and cortical (Figure 5.11) labelled image regions. However, the variational Bayesian approaches produce smoother registration transformations (Figure 5.12), with less image folding (Figure 5.13). Furthermore, for certain registrations in the IBSR dataset, the proposed approaches were shown to be more accurate, an example

illustrating this is given in Figure 5.14, where a higher level of regularisation was required. The local noise model was also shown to be more effective than a global noise model for accurately registering some structures, as shown in Figure 5.15.

The next chapter explores the concept of uncertainty in non-rigid registration and describes how estimates of registration uncertainty, such as those produced by the probabilistic registration frameworks, can be used to compensate for residual mis-registration. An adaptive smoothing filter based on the registration uncertainty is proposed, and is demonstrated to improve anatomical correspondence, as measured by subcortical segmentation labels.

Chapter 6

Registration Derived Uncertainty

6.1 Introduction

In this chapter the concept of uncertainty in medical image registration is explored. The interpretation and visualisation of the uncertainty estimated from the probabilistic registration algorithms is illustrated. An adaptive local smoothing filter is introduced, which is derived from the estimated registration uncertainty. This approach provides a computationally tractable method to compensate for registration uncertainty in high resolution, whole brain image registration. Smoothing of spatially normalised data is motivated, and the benefits of the adaptive smoothing filter are explored using subcortical segmentations. The uncertainty derived smoothing filter is shown to provide a better trade-off of sensitivity and specificity than Gaussian filtering. This work is based on previous conference publications [165][164].

6.1.1 Motivation

Medical image registration has a great deal of intrinsic uncertainty associated with the inference of an optimal mapping between images. Structural MR images of the brain only provide a noisy measurement of the tissue properties of the underlying anatomy. Therefore, any intensity based approaches to image registration will naturally contain ambiguities. This is because the intensities in a given location are not unique. This can be somewhat alleviated by additionally using image derived features, for example [162], or more effectively by mutual saliency [132].

However, this only provides accurate cues around certain image features, such as edges, which are usually well aligned anyway.

The ambiguity of corresponding anatomy in brain MRI means that for the problem of inter-subject registration, the image data is insufficient to support a single "true" solution. The registration problem is further confounded by anatomical variability, and the possibility that certain sulci may not have homologues in every subject. As discussed in Chapter 5, in a large comparison study, Klein et al. [105] found that none of the registration algorithms currently available were able to produce a perfect mapping between subjects, as measured by the overlap of anatomical segmentation labels. As no single mapping perfectly describes the relationship between images, the set of likely mappings should be investigated.

The distribution of probable mappings is derived from the ambiguity of image matching. Boundaries between two intensity regions in an image provide a smaller range of potential matching points than in homogeneous regions. Nevertheless, there will still be a degree of uncertainty in the exact matching position along a boundary, albeit with much less across it. Additionally, one would expect any probable mapping between images to be smooth, so a spatial prior should be included in the uncertainty. This provides additional information to avoid estimating every point in a homogeneous region as equivalent. As such, the trade-off between data fidelity and spatial prior is important for understanding the uncertainty of registration. The ambiguity, or residual mis-registration, will affect any conclusions that are drawn from registered medical images. For this reason, in many neuroimaging applications, spatially normalised data is smoothed using a Gaussian kernel to compensate for mis-registration. This approach requires the heuristic definition of the size of an appropriate smoothing kernel, which may not be optimal for the particular data.

6.1.2 Previous Work in Registration Uncertainty

Previous work on estimating the uncertainty in non-rigid registration includes that of Hub et al. [93]. They stochastically estimate the variability of a chosen cost function with respect to the transformation parameters. This allows the estimation of regions of uncertainty in the mapping due to insufficiently discriminative information available in the data. This approach is generic in that any cost function could be used, including statistical measures such as mutual information. However, they neglect to include estimation of the regularisation cost in their uncertainty, which will undoubtedly affect the probability of a mapping.

More principled work on estimating uncertainty includes that of Allassonnière et al. [2]. They describe a Bayesian deformable template registration framework. Their approach models individual images as random deformation of an estimated template, assuming a fixed prior on the deformations from the template. This model is inferred upon using expectation-maximisation (EM), or a stochastic approximation to EM [3]. Their inference strategy yields an approximate posterior distribution of transformation parameters, which may be modelled as a mixture of multi-variate normals. Their work has the limitation that the strength of the prior, which greatly affects the posterior distribution, has to be hand-defined. This algorithm was demonstrated on 2D digit recognition.

An alternative view of registration uncertainty has been proposed by Van Leemput [183], which was previously described in section 3.1.2. Their approach does not attempt to estimate the distribution of the true deformation field, but rather models the distribution of an image segmentation labelling. Such a method may provide complementary uncertainty information to estimating the distribution of mappings. Unfortunately, no 3-D results are presented due to the computationally complexity of sampling the posterior distribution.

Probabilistic sampling methods that numerically estimate the transformation parameter distribution have been proposed [67][64][146]. Gee et al. [67][64] used a Gibbs sampler to estimate the mean of the posterior distribution, and they report estimates of the variance of transformation parameters in 2D registration. Risholm et al. [146] describe a a Markov chain Monte Carlo based approach to marginalise over model hyper-parameters in low resolution 3D registration [147]. This allows an estimate of the registration uncertainty without making distributional assumptions about the posterior, in contrast to the approach in this thesis. For sufficient samples, sampling gives the best estimate of the true transformation parameter distribution. However, due to the associated computational complexity of numerical integration of probability distributions, such an approach can only be used with limited degrees of freedom. Risholm has investigated approaches to visualise the uncertainty distribution [147], estimating the uncertainty in prostate intervention [145] and the uncertainty in dose delivery in radiotherapy [143]. However, their use of such a coarse transformation model, and naïve application of SSD without an intensity mapping, means that the uncertainty in registration appears to substantially larger than would be expected.

6.1.3 Proposed Solution

The contribution of this Chapter lies in the derivation and interpretation of a voxelwise measure of registration uncertainty. From this basis, it is demonstrated how a spatially varying smoothing filter can be estimated and used to compensate for registration uncertainty. This information is derived from the registration framework that was proposed in Chapter 3. The use of variational Bayesian inference provides an intrinsic estimate of the uncertainty of the model parameters. Of these parameters, most interestingly there is an estimate of the uncertainty of the transformation parameters. This takes the form of a covariance matrix $\boldsymbol{\Upsilon}^{-1}$. The uncertainty of transformation parameters provides a measure of the uncertainty of the deformation field. Therefore, it also describes the uncertainty of the voxel intensity in the transformed source image, and of any registration derived features, such as Jacobian maps. The registration framework is computationally tractable for high resolution registration, and provides an estimation of the level of spatial and noise precision, the values of which are important for uncertainty estimation.

$\boldsymbol{\Upsilon}^{-1}$ can be interpolated to the voxel level, providing an estimate of voxelwise spatial variance and directional covariance. This chapter describes how these voxel level estimates of spatial uncertainty can be used to estimate an adaptive local smoothing filter, which helps compensate for residual mis-registration.

6.2 Methods

This chapter uses the fully Bayesian probabilistic registration frameworks described in Chapters 3 and 4, which are inferred upon using variational Bayes. The uncertainty of model parameters are encapsulated into parametric probability distributions, and the posterior distribution of transformation parameters is inferred as a multivariate Normal distribution. These parameter distribution estimates are subject to the 1st order Taylor series expansion on the transformation function, as discussed in section 3.3.3. This makes the assumption that the transformed image varies locally linearly, with respect to the transformation parameters, $\mathbf{w}$. Furthermore, the approximate posterior distributions of the transformation, noise and regularisation parameters are approximated as independent, as discussed in section 3.3.1.

The uncertainty of $\mathbf{w}$ is encoded in the covariance matrix $\boldsymbol{\Upsilon}^{-1}$. This is stored as a precision matrix, as it is sparsely populated, with a non-zero structure dic-

tated by the order of the B-splines, and the form of the regularisation. The covariance form is much less sparse. In whole brain, high-resolution registration, with a $5mm$ b-spline knot spacing, the number of degrees of freedom, N_c, would be approximately 180,000. Therefore, the covariance matrix, if full, would require at least 120 Gigabytes of RAM to store, making a matrix inversion computationally intractable. As such, it is not possible to draw samples from the full distribution of probable transformations in whole brain registration, although this is later explored for region of interest (ROI) analysis in Chapter 8. However, such analysis is limited in requiring pre-selection of the ROI. This chapter investigates an approximate and tractable application of uncertainty in whole brain analysis.

6.2.1 Spatial Uncertainty

The uncertainty of the transformation is given by $\boldsymbol{\Upsilon}^{-1}$, which has units of mm^2. As any change in the value of the transformation parameters, leads to a change in the deformation field, the uncertainty of $\mathbf{w}$ corresponds to the uncertainty of the estimated voxel to voxel mapping.

As the uncertainty of $\mathbf{w}$ is stored in a precision form in $\boldsymbol{\Upsilon}$, it needs to be converted to a covariance matrix to obtain the variance and cross-directional covariance. In this thesis only the co-precision terms between control point directions are included for the purposes of inversion. A discussion of this approach is given in section 6.4.1. The resulting spatial uncertainty information for a particular transformation parameter, as described in equation 3.14, is governed by four factors:

- The local image information that is affected by any changes in the parameter value from the current mean estimate, $\mathbf{J}^T\mathbf{J}$.
- The noise precision: how much the image data is trusted, related to the sum of squared differences of the residual, corrected for spatial smoothness, $\alpha\phi$.
- The form of the spatial prior: e.g. bending energy, membrane energy, Λ.
- The spatial precision: how similar the transformation is to the spatial prior, λ.

The estimated variance and cross-directional covariance provide sufficient information to describe an independent multivariate Normal distribution for each transformation parameter. For a B-spline FFD transformation model, this can

be propagated to the voxel level by interpolating the variance and covariance of each transformation parameter in each direction, using the B-spline basis set. This allows the estimation of a multivariate Normal distribution that describes the spatial uncertainty at each individual voxel. This distribution describes the magnitude and direction of the voxelwise spatial uncertainty. As the uncertainty measure is dependent on the image information, it is lower across an image boundary than along it. This results in an anisotropic measure of spatial uncertainty. The scale of uncertainty will vary across individual registrations, depending on the inferred level of noise and spatial precision.

In the analysis of medical images, feature data, such as longitudinal morphometric features or anatomical segmentations, may need to be transformed from one image domain to another. The transformation between image domains is estimated by registration, which will be uncertain. An approach to compensate for the spatial uncertainty in the mapping of any given voxel, is to smooth the transformed feature data according to the local uncertainty distribution. As a voxelwise anisotropic Gaussian distribution can be calculated for each voxel, this can be used as a smoothing kernel. This allows the calculation of an uncertainty compensated feature value at each voxel. This scheme is proposed to reduce the effects of ambiguous matching and residual mis-registration in feature data that has been transformed to a different image space.

6.3 Results

6.3.1 Overview of Experiments

The 306 registrations of the subject pairs in the IBSR dataset, which was used in the previous chapter, were used in all of these experiments to investigate the effects of registration uncertainty. This section begins with visualisation of the average, and variability in the estimated spatial variance of registration. The effects of using a local or global noise model is also investigated. This is followed by an example of a more detailed visualisation of the uncertainty distribution.

The subsequent experiments focus on the effects of smoothing propagated anatomical labels. This provides a probabilistic view of propagated segmentations. The benefits of smoothing segmentations using the registration derived uncertainty are examined in terms of a binary classifier, and compared against a standard Gaussian smoothing approach. A more general description of the use of smoothing in spatially normalised analysis is reserved for section 7.1.2.

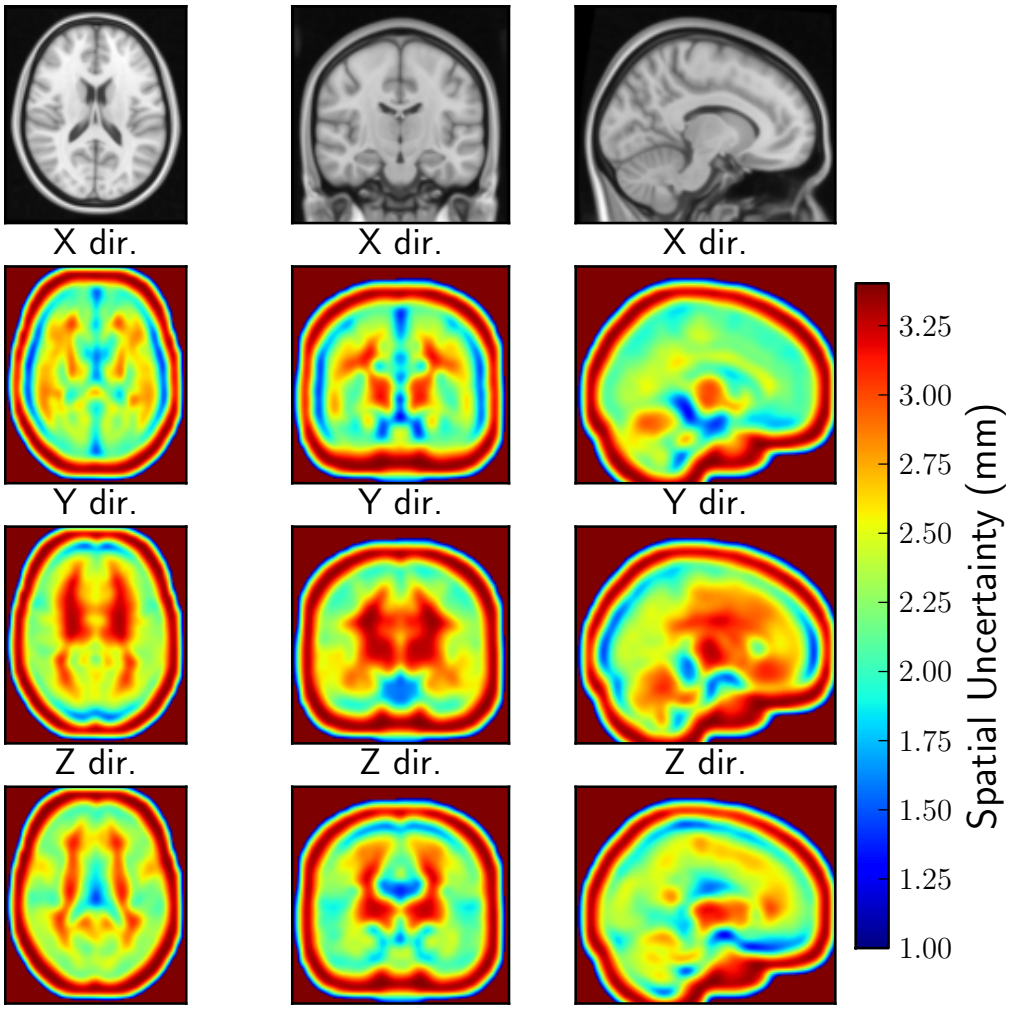

Figure 6.1: Map of the average spatial variance of the inferred transformations from the 306 pairwise IBSR registration using the global noise model. The top row shows the MNI152 atlas, which each of the images was originally affinely aligned to. The next three rows show the average voxelwise registration uncertainty for each transformation direction. X corresponds to left-to-right, Y posterior-to-anterior and Z inferior-to-superior. Notice the lower variance surrounding edges, and higher uncertainty in homogeneous regions.

6.3.2 Visualisation of Uncertainty

Visualisation of Voxelwise Spatial Variance

To assess the estimated uncertainty, it is useful to be able to visualise it across the brain. The simplest depiction of uncertainty is showing the variance in each direction. A map illustrating the average estimated spatial variance in each direction

is given in Figure 6.1. This information is averaged across the 306 registrations from the IBSR dataset with a global noise model. The difference in the average level of uncertainty estimated using a global, or local noise model is plotted in Figure 6.2. The effects of using a local noise model appears to be quite small.

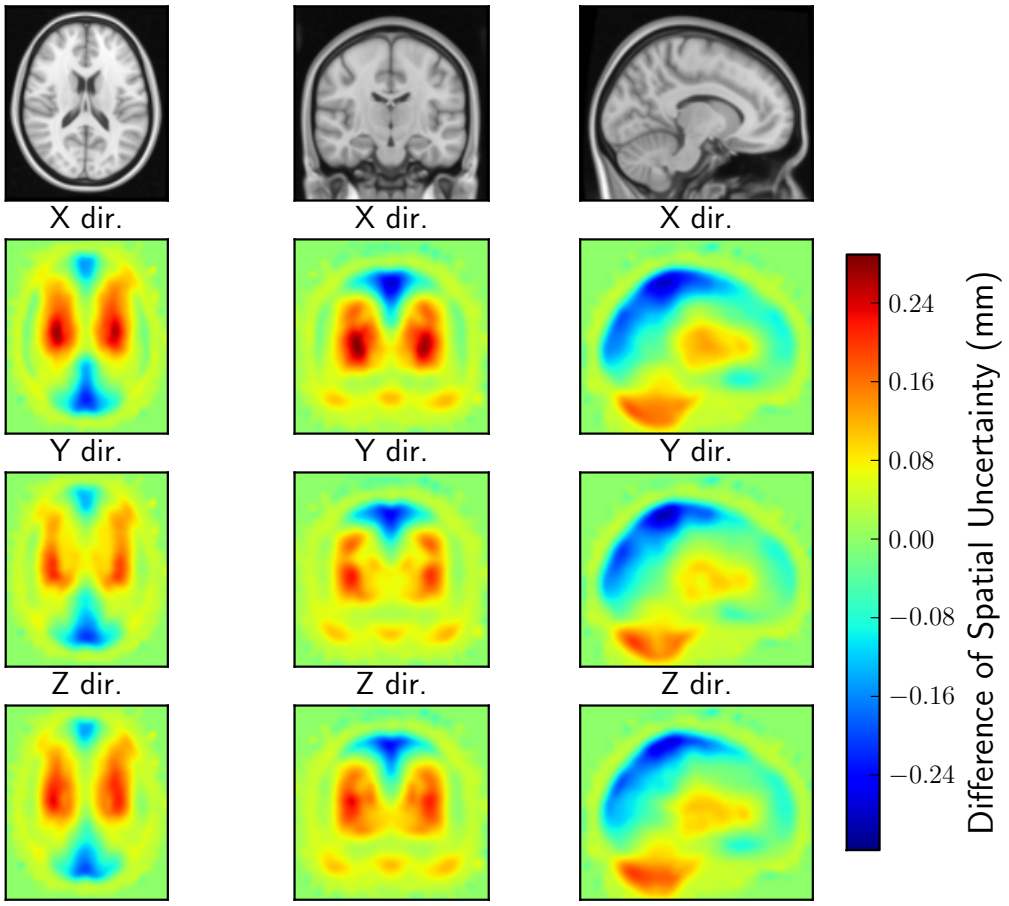

Figure 6.2: Map of the difference in average spatial variance of the inferred transformations from the 306 IBSR registration using the global or local noise model. The top row shows the MNI152 atlas, the next three rows show the average local noise variance subtracted from the average global noise uncertainty. Using a local noise model results in lower uncertainty in white matter regions, and an increase in uncertainty in cortical regions. The difference in uncertainty is consistent across directions, this is because the image derived uncertainty component maintains the same direction, but is scaled differently across the brain.

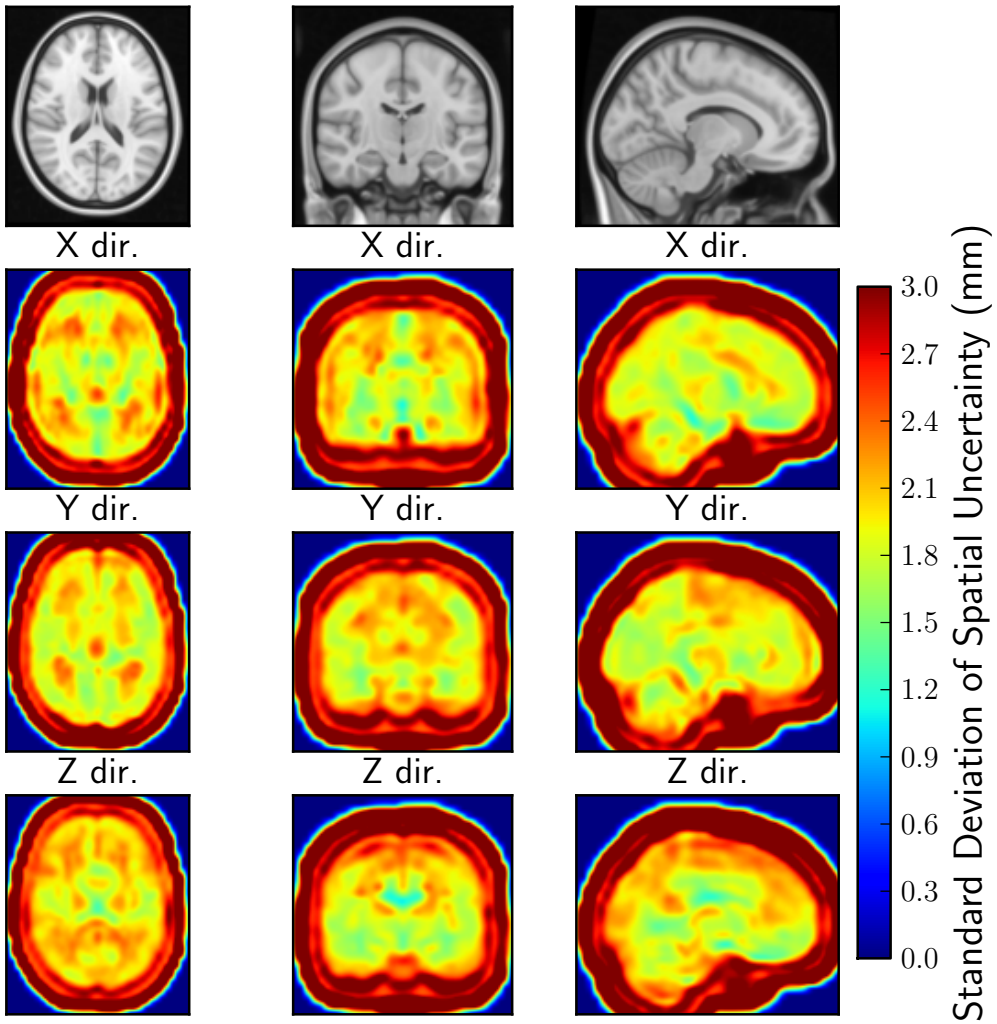

Figure 6.3: Map of the standard deviation of spatial variance in the inferred transformations across the 306 IBSR registration using the global noise model. The top row shows the MNI152 atlas, the next three rows show the standard deviation of the voxelwise registration uncertainty for each transformation direction. There appears to be a large degree of variation across subjects.

Visualisation of Variability in Voxelwise Spatial Variance

The voxelwise variability in variance across subjects is interesting as it shows how data dependent the uncertainty information is. This information is plotted for the global noise model in Figure 6.3. A high degree of variability in uncertainty across subjects is visible. This is more visible at strong edges because the uncertainty is almost entirely dependent on the image data here, so a difference in noise precision makes a big difference. The difference in voxelwise variability in spatial

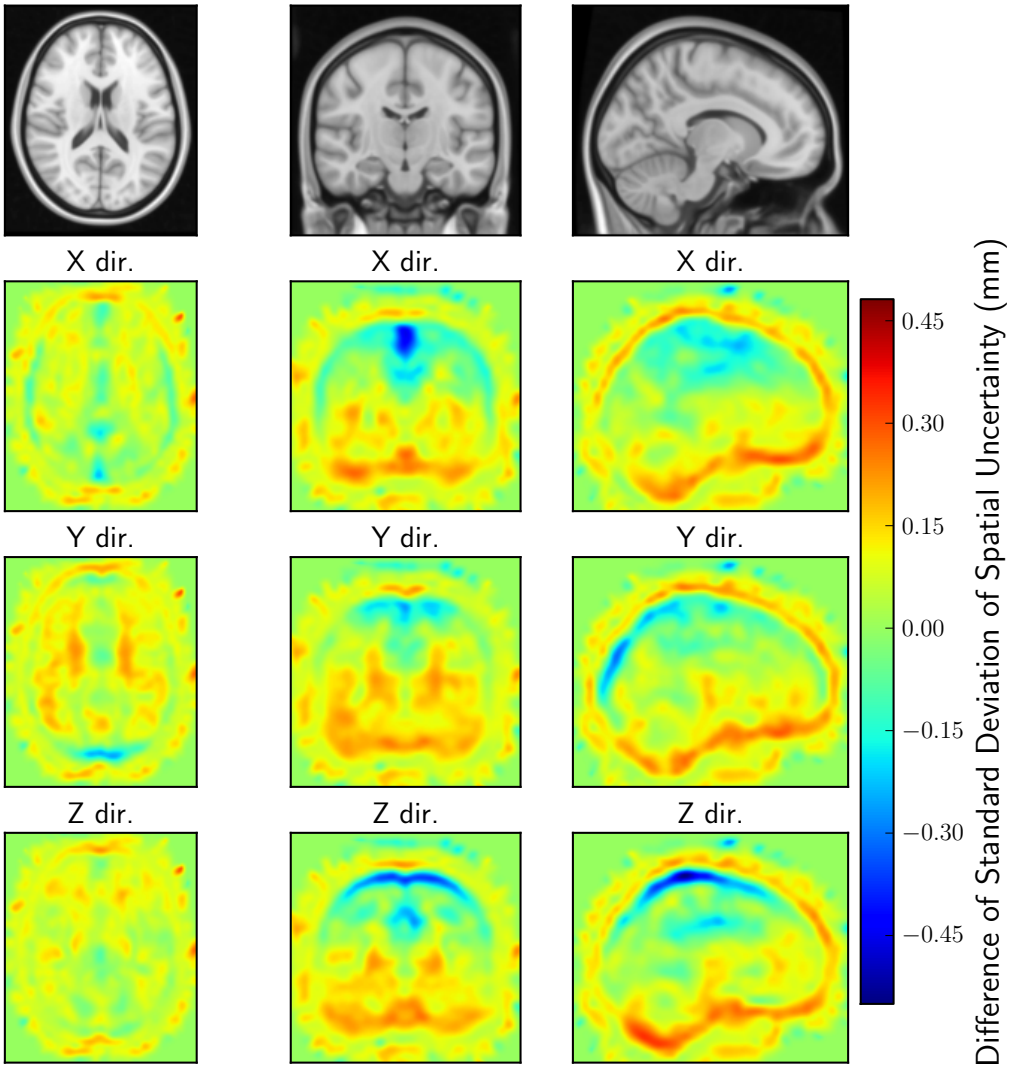

Figure 6.4: Map of the difference in standard deviation of spatial variance of the inferred transformations from the 306 IBSR registration using the global or local noise model. The top row shows the MNI152 atlas, the next three rows show the standard deviation of the spatial variance from the local noise model subtracted from the global model. Using a local noise model leads to more variability in uncertainty of cortical regions and less in homogeneous regions and at the base of the brain.

variance across subjects when using either a local or global noise model is shown in Figure 6.4.

Visualisation of Voxelwise Uncertainty Distribution

An alternative visualisation strategy that includes the covariance between directions is to plot ellipsoids to illustrate the voxelwise uncertainty. An example is given in Figure 6.5. The axes of each ellipsoid represents the full-width at half maximum of the uncertainty distribution.

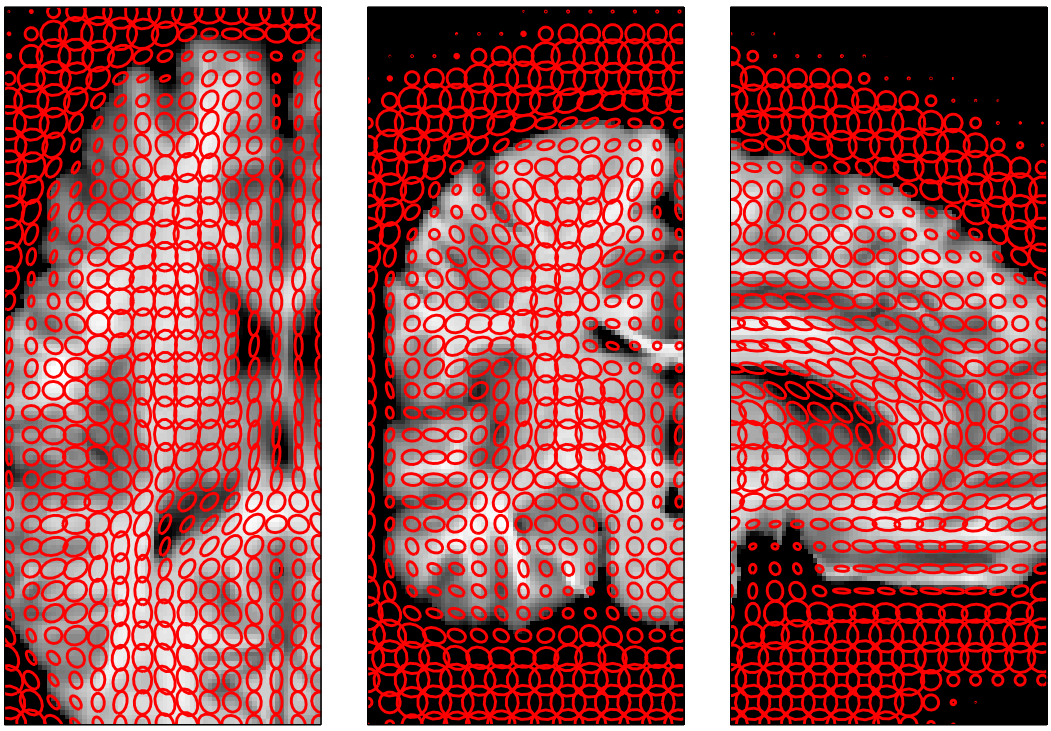

Figure 6.5: An example plot of the voxelwise uncertainty distribution overlaid on the transformed subject image. Each ellipse represents the full-width at half-maximum of the uncertainty distribution in that plane at that location. The ellipse centres are plotted 5mm apart, and the registration used a 5mm B-spline knot spacing. Note the directional component adjacent to strong edges.

6.3.3 Segmentation Propagation

The principal motivation for smoothing transformed feature data, is to maximise the amount of image information in a voxel in the reference space, which comes

from the true corresponding voxel in the registered image. As voxelwise correspondence is unknown, anatomical segmentations can again be used as a surrogate measure of correspondence. In these experiments, the subcortical anatomical labels in the IBSR dataset were individually transformed from the source image space, to the target image space using the mean of the transformation parameters, μ. These inter-subject registrations mimic the spatial normalisation problem. Once transformed, each label can be smoothed, as is commonly applied to spatially normalised data, which provides an estimate of the posterior distribution of the segmentation. Subsequently, the benefits of using a registration uncertainty derived smoothing method can be evaluated, and compared against traditional isotropic Gaussian smoothing.

Example Probabilistic Segmentation Label Propagation

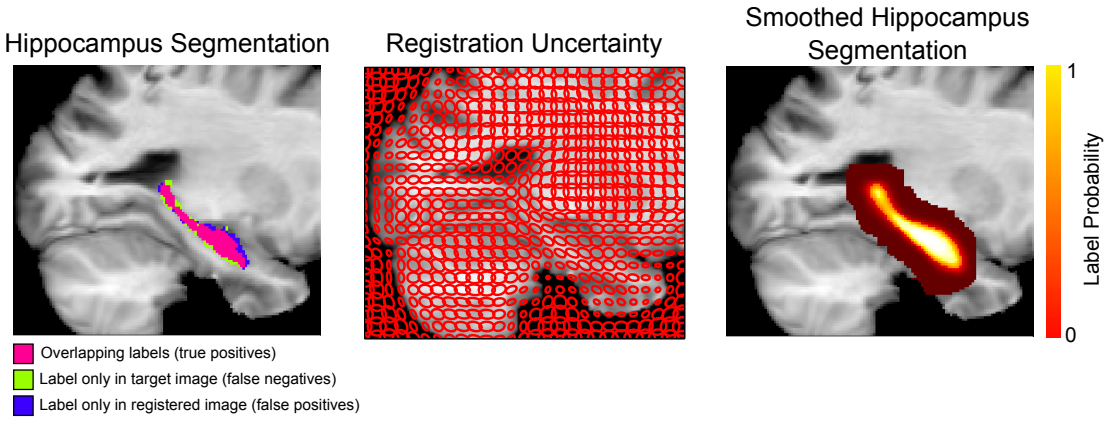

Figure 6.6: An example of segmentation label propagation, and the application of uncertainty based smoothing for the right hippocampus. These images were registered using the VB registration algorithm with a global noise model. The left image illustrates the overlap of the propagated segmentation label. The middle panel illustrates the registration uncertainty for this slice. The right image shows the label probability map that arises from smoothing the transformed label by the registration uncertainty. The directionality of the uncertainty means the label is less smoothed across image edges than with a spherical Gaussian kernel.

An example of segmentation label propagation, and smoothing to compensate for registration uncertainty is given in Figure 6.6. This shows a non-zeros probability is given to every voxel in the true label location, as well as to many that are not. To produce a hard segmentation, an appropriate probability threshold would need to be selected.

6.3.4 ROC analysis

A different mechanism to those described in the previous chapter, are required to evaluate probabilistic propagated anatomical segmentations. Previous work in medical image segmentation has investigated the use of receiver operating characteristic (ROC) curves to evaluate probabilistic segmentation accuracy [210]. ROC curves treat the segmentation problem as a binary classifier [123]. In this case, the classifier output is whether a voxel contains the anatomical label of interest. The ROC curve is constructed by varying the probability threshold to create several hard segmentations. This describes the effect of the threshold on the classifier performance. The performance can be separated into measures of the sensitivity and specificity of the estimated segmentation with respect to the true segmentation. Sensitivity, otherwise referred to as the true positive rate (TPR), is defined as:

$$TPR = \frac{TP}{TP + FN} \tag{6.1}$$

where TP is the count of true positives and FN is the count of false negatives. The false positive rate (FPR), which is equivalent to 1-specificity, is defined as:

$$FPR = \frac{FP}{(FP + TN)} \tag{6.2}$$

where FP is the count of false positives, and TN is the count of true negatives. ROC curves plot TPR against FPR. In this chapter, ROC curves are used to evaluate the performance of different image smoothing methods on the segmentation of anatomical structures. Any smoothing method will increase the sensitivity of the segmentation, as more voxels are considered part of the structure. However, this may lead to an increase in false positives. An example ROC curve is given in Figure 6.7.

ROC Curves for Segmentation

A difficulty in performing a ROC analysis of probabilistic medical image segmentations is that each segmentation label is evaluated as a separate binary classifier. Depending on the size of the anatomical structure, and the image, this can lead to a very large amount of true negatives, which dwarf the false positive count. This can result in the number of false positives being under-represented in the false positive rate. As a solution to this problem, the count of background voxels (FP+TN) for each segmentation label type was specified manually. The count of

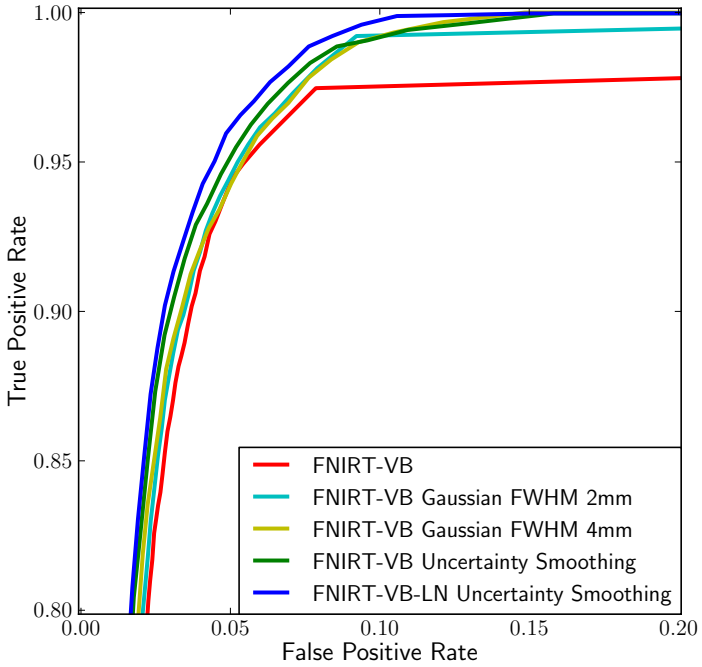

Figure 6.7: Example ROC curves from right hippocampus segmentation using different image smoothing strategies for the same registration as shown in Figure 6.6. The classifier that provides the highest true positive rate, for the lowest false positive rate is the most effective. In this example, smoothing the data using the registration uncertainty inferred from local noise image provides the best classification. Using the uncertainty from the global noise model is more effective than either Gaussian smoothing approach. Not smoothing the data leads to lower sensitivity.

background voxels for a given anatomical structure was selected to be the largest number of false positives given by any of the smoothing methods on any of the registrations. This ensures a more reasonable weighting of FP in the FPR. This also provides a fair test in that all ROC curves for a particular segmentation label, are based on the same number of background voxels.

ROC curves are commonly used to estimate the ideal threshold for a binary classifier. However, in these experiments the optimal threshold will depend on how accurately the structure was localised, and thus will differ between registrations. Moreover, the objective of this work is not to provide a method for generating the most accurate hard segmentations, but instead to find the smoothing approach that best compensates for mis-registration in inter-subject registration. The ideal smoothing method will provide the optimal sensitivity to specificity trade-off for each label, across the set of images. This means that the image

information in any voxel, in a given anatomical structure of the reference image space, will be maximally sensitive to the image information in the corresponding structure of the registered source image. Furthermore, the image information at this voxel will be more specifically drawn from the corresponding anatomical region than any other.

Area Under the Curve

For the purpose of evaluating the trade off of sensitivity and specificity of a binary classifier, the area under the ROC curve (AUC) [175], provides a useful summary statistic. The AUC statistic represents the probability that, for a random observation of a voxel that is part of the true segmented region and a random background voxel, the results will be ranked correctly in terms of the estimated class probability. The AUC has been shown to be related to to the Mann-Witney U statistic [20], and has been used to perform comparisons between multiple models for individual ROC curves [77][47][28]. In this work there are 306 ROC curves for each segmentation label. Therefore, it is more convenient to present a boxplot of the distribution of AUC statistics for each segmentation label. This allows the effects of different smoothing methods to be compared.

In this experiment, the uncertainty derived smoothing method using a global or local noise model, is compared against Gaussian smoothing. Three different Gaussian kernel smoothing kernels were tested, with a full-width at half-maximum (FWHM) of 2mm, 4mm and 8mm. The results using a FWHM of 8mm were consistently worse than all other approaches, and thus are not presented. The AUC of each ROC curve is calculated by trapezoidal approximation. The boxplot of AUC statistics is presented in Figure 6.8. The uncertainty derived smoothing performs at least as well, but usually better than either Gaussian smoothing approach for all labels except the putamen, cerebral cortex and white matter. The image information is weak surrounding the putamen, thus this structure may be over-smoothed. The difference in performance in the cerebral cortex and white matter regions is probably due to the uncertainty information being too smooth to account for the geometrically complex structures. Additionally, the cerebral cortex and white matter labels cover a very large area, which has disparate functionality. For the grey matter it would be much more appropriate to consider cortical labels, if accurate segmentations were available. This has not been investigated in this work.

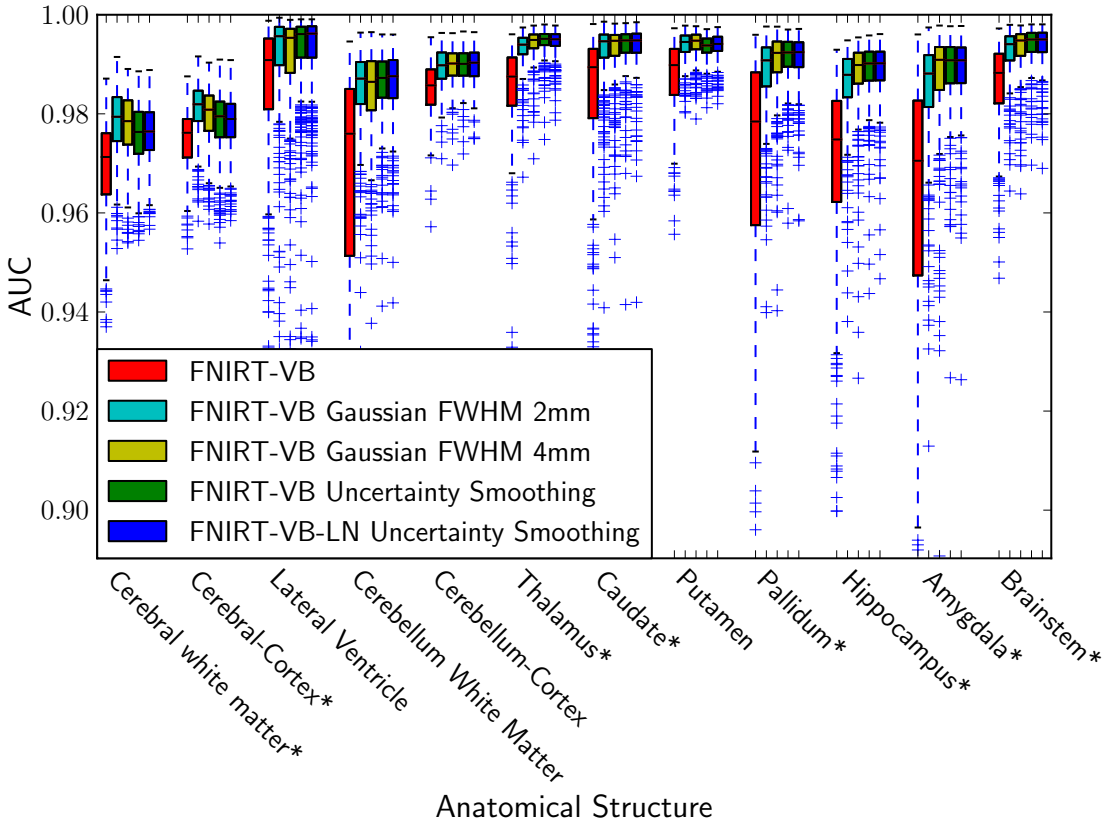

Figure 6.8: Boxplot of the area under the curve (AUC) of the ROC curves for each segmentation, under different smoothing approaches. The * at the end of a region name indicates a statistically significant difference in target overlap between the 4 smoothing methods, as measured by one-way analysis of variations (ANOVA), $p < 0.05$). All smoothing methods provides an improvement in AUC over not smoothing. Some structures, such as the thalamus and hippocampus show slight improvements using uncertainty smoothing in the average AUC compared to either Gaussian smoothing method. Apart from the cerebal cortex, and white matter, the remaining structures seem to do similarly well on average using either uncertainty derived smoothing or Gaussian smoothing. Interestingly, the lower quartiles and outliers seem to be improved when using uncertainty derived smoothing for several structures, including the amygdala, cerebellum white matter, caudate and brainstem. There seems to be little difference in the affects of uncertainty based smoothing using either a global or local noise model.

6.4 Discussion and Conclusions

6.4.1 Discussion

Smoothing of spatially normalised data has been motivated in terms of improving the trade-off of sensitivity and specificity of anatomical segmentations. This

chapter has demonstrated that smoothing anatomical segmentations improves the trade-off of sensitivity and specificity of the propagated label to the true segmentations. Furthermore, the use of a smoothing method derived from registration uncertainty was shown to outperform Gaussian smoothing for most subcortical structures.

A limitation with the registration uncertainty calculated using either of the presented VB registration frameworks lies in the mean field approximation, $q(\mathbf{w}, \phi) \sim q(\mathbf{w})q(\phi)$. This means that the joint distribution of the model noise with the transformation parameters is not calculated. Consequently, the image information incorporated into $\boldsymbol{\Upsilon}$ from $\mathbf{J}^T\mathbf{J}$ is necessarily drawn from the source image only. Furthermore, the ambiguity of intensity matching in the target image will follow a different shape to that in the source, unless the two images are perfectly aligned. Therefore, a more comprehensive model of uncertainty should incorporate this, and this is discussed in Chapter 9.

From the results in this, and the previous chapter, the use of a local noise model does not seem to produce a substantial effect. This can be attributed to the use of a smooth basis set, with limited degrees of freedom, smoothing out the effects of varying noise precision. As insufficient improvement is found, and the computational cost is significantly increased, the experiments in Chapters 7 and 8 will only consider the global noise model.

As the uncertainty of $\mathbf{w}$ is stored in a precision form in $\boldsymbol{\Upsilon}$, it needs to be converted to a covariance matrix to be used. The only entries of the covariance matrix that are required are the variance, and cross directional covariance of the transformation parameters. This leaves a choice of whether to calculate these terms by including the covariance between transformation parameters, or not. The covariance between control points would be essential if transformations were being sampled to ensure the transformations are smooth. However, including these entries in the inversion leads to the estimation of larger variances. As the covariance between control points is not used, it does not seem appropriate to calculate the variance and cross directional covariance terms to include this.

In this work the strength of a prior with a fixed covariance structure, based on bending energy, is inferred from the data. A fixed prior covariance structure expects a similarly smooth transformation across the image. The choice of structure will effect the posterior transformation distribution, particularly in regions where little image information is available. The structure of this prior distribution could be estimated from the data. Such an approach could provide a more biologically plausible prior. This is further discussed in Chapter 9.

If the objective was the accurate segmentation of anatomical structures, then an approach to uncertainty derived smoothing may still be beneficial. The simplest approach would be to smooth every label, and then take the most probable label for every voxel. However, this would probably have the effect of shrinking smaller structures. Alternatively, registration uncertainty could be incorporated into a label fusion framework to allow accurate segmentations of new subjects [152][189].

An alternative strategy to compensate for spatial uncertainty would be to build up a voxelwise distribution of intensities according to the registration uncertainty. This can be achieved by sampling with the relevant frequency from surrounding image voxels. Such an approach could improve the level of spatial resolution of an analysis as voxels between different populations could be treated as mixtures, rather than by smoothing across the populations. This is not investigated in this thesis as a distribution would need to be analysed at each voxel, instead of a single value.

6.4.2 Conclusions

This chapter has introduced approaches for visualising registration uncertainty, and provided some illustrations of the level of uncertainty calculated by the probabilistic registration algorithms proposed in this thesis. Subsequently, a generic, principled mechanism for compensating for the uncertainty in non-rigid registration was introduced in the form of a smoothing kernel. The use of data smoothing is explored in the context of anatomical segmentation propagation. Registration uncertainty derived smoothing is demonstrated to provide an effective approach to compensate for residual mis-registration.

The next chapter explores the application of the probabilistic registration framework to derive longitudinal morphometric features of Alzhiemer's disease (AD). These are spatially normalised to a representative atlas. The uncertainty of the spatial normalisation is compensated for by smoothing the transformed longitudinal feature maps, and the uncertainty derived filter is compared against Gaussian smoothing. The smoothed feature data is used for the statistical prediction of a subjects' disease status, and is shown to be highly discriminative between AD and controls.

Chapter 7

Longitudinal Analysis Of Alzheimer's Disease

7.1 Introduction

This chapter describes an approach for the longitudinal analysis of Alzheimer's disease (AD), with the aim of best discriminating subjects suffering from AD from age matched normal controls (NC). Discriminative longitudinal features are generated through the use of the Bayesian probabilistic non-rigid registration algorithm. To perform statistical analysis on these features, they are spatially normalised through the registration of the baseline images to a reference space. The benefits of smoothing the data using the registration uncertainty derived kernel, which was proposed in the previous chapter, are further demonstrated in the subsequent spatially normalised statistical analysis. A previous version of this work has been presented [164].

7.1.1 Motivation

As described in sections 1.2.1 and 2.3.2, longitudinal structural MR imaging of subjects with AD can be used to provide an objective measure of localised anatomical changes, such as those found in the progression of neurodegenerative disease [59]. Non-rigid registration methods can be used to describe longitudinal changes in a subject. This is achieved through the analysis of the deformation field tensor in a framework known as tensor based morphometry (TBM) [38].

A difficulty in estimating a reasonable set of features lies in the selection of the level of regularisation. As no ground truth schemes exist, it is difficult to find an optimal level of regularisation without manual intervention. As described in section 2.3.2, two surrogate measures of TBM feature accuracy have been proposed, but these have limitations and provide no guarantee as to the accuracy of TBM features from real data.

To establish a consistent set of anatomical changes, statistical analysis of longitudinal TBM features needs to be carried out across a population. This necessitates the spatial normalisation of all subjects to a common reference space. This requires inter-subject registration, which has been shown in the previous two chapters to be an inexact and uncertain process. The most common approach to compensate for mis-registration is through Gaussian smoothing of the data. A disadvantage of such an approach is an appropriate sized Gaussian kernel needs to be chosen, which may not be optimal for the whole image.

7.1.2 Previous Work

Longitudinal TBM

While there are many studies of longitudinal volumetry, or whole brain atrophy rate estimation in Alzheimer's disease (see [158] for a review), there are relatively few studies that have been proposed for the analysis of longitudinal TBM. Scahill et al. [159] estimate longitudinal volume change using a fluid registration algorithm [59], where the parameters of the method were optimised through simulated hippocampal atrophy [43]. The log of the Jacobian was calculated, and the feature maps were split into an image of expansion, and one of contraction. This was to prevent mixing the populations after smoothing. The TBM features were spatially normalised, and smoothed using a Gaussian kernel with a FWHM of 8mm. The statistical differences between the two groups were analysed in SPM99, which uses a Bayesian affine registration [15] and global basis set non-rigid registration [10] to perform the spatial normalisation. Maps of voxelwise statistical significance were analysed for differences between two populations.

Leow et al [112] and Hua et al. [88] investigated longitudinal changes in subjects with AD, mild cognitive impairment and controls. This data was taken from the Alzheimer's disease neuroimaging initiative [129]. Longitudinal images were registered using an unbiased fluid registration algorithm [111], with a mutual information cost function, which performed well in [200]. The spatial normalisation used an elastic non-rigid image registration algorithm [110], which also

used a mutual information cost function. Smoothing of feature data was not reported. Voxels were correlated with clinical measurements, and statistical differences between subject groups were analysed. Hua et al. [88] also investigated the advantage of using longer or shorter scan intervals. The predictive capacity of the data was not investigated for either of these studies.

Smoothing of Spatially Normalised Data

There is limited work that discusses the rationale, or alternatives to data smoothing in spatially normalised analysis. Worsley et al. [197] state that smoothing spatially normalised data serves the purpose of compensating for mis-registration. They also suggest that where possible, the selected size of the smoothing kernel should be related to the size of the region of interest in the image. The rationale for this comes from the matched filter theorem [155], which states that: the signal to noise ratio of the signal of interest can be optimally boosted by using a filter of the same size and shape as the signal of interest. Ashburner and Friston [11] also note that data smoothing leads to the data becoming more Normally distributed, which increases the validity of certain analysis approaches. Ridgway [141] provides a qualitative exploration of several smoothing kernel sizes in log-Jacobian TBM and deformation based morphometry. He found that an 8mm smoothing kernel provided the most visually appealing results, smaller kernels were found to produce less well connected blobs, and larger kernels over-smoothed the data. However, any such results are likely to be quite dependent on the data being processing, and the registration procedure used.

For certain generative statistical analysis models, Bayesian approaches have been designed that permit the inference of the level of isotropic spatial smoothness from the data features of interest [135]. These have been extended to allow an anisotropic diffusion smoothness estimation [78]. Such an approach is computationally very expensive, and so approximate solutions that assume regional independence are sometimes used [79]. The only other work that uses estimates of registration uncertainty to improve spatially normalised analysis is Keller et al. [103]. They propose a hierarchical Bayesian framework where an approximate model of registration uncertainty is integrated out of a voxelwise decision statistic. They noted their approach reduced the artefactual "stretching effects" that can occur under Gaussian smoothing. However, their approach to modelling registration uncertainty is very simple, using an approximation of a fixed spherical Gaussian.

7.1.3 Proposed Solution

This chapter demonstrates the benefits of using a fully Bayesian approach to non-rigid registration for analysing structural changes in the brain of subjects with Alzheimer's disease, from longitudinal MR images. Firstly, it is shown how the probabilistic non-rigid registration framework can be applied to derive informative features from longitudinal data. Subsequently, the advantages of using a probabilistic approach to registration are highlighted in the spatial normalisation of the subject data. The registration uncertainty derived smoothing filter that was introduced in the previous chapter, is used as a means of compensating for the inherent mis-registration involved in spatially normalised statistical analysis. As this approach is an alternative method of data pre-processing, it can fit into any analysis framework. The discriminative capabilities of the data within this framework are illustrated in terms of voxelwise statistics, and subject disease status prediction using multi-variate classifiers.

7.2 Materials

7.2.1 ADNI

The data used in this, and the following chapter was obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database [129][1]. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as to lessen the time and cost of clinical trials.

ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to

[1] adni.loni.ucla.edu

Table 7.1: Statistics of the training and testing subject groups.

Subject Group	Mean Age (years)	Age Standard Deviation (years)	Mean MMSE
AD Train	77.4	7.47	19.4
NC Train	78.7	5.57	29.0
AD Test	76.4	7.50	20.4
NC Test	78.9	4.82	29.1

recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

7.2.2 Subject Grouping

The objective of this chapter is to explore the predictive capabilities of longitudinal TBM data with respect to the presence of AD. The experiments need to be designed to reflect the real world situation, where unknown data, rather than previously observed examples are used to measure the performance. This necessitates the separation of data into training and testing sets. To ensure a fair test, the testing data and its labelling must remain completely unknown in the training phase.

In this work a total of 311 subject images were taken from the ADNI database. 149 of these subjects were patients with Alzheimer's Disease (AD), and 162 were age matched normal controls (NC). These images were broken up into training (81 AD, 81 NC) and testing (68 AD, 81 NC) sets. The properties of these groups are given in Table 7.1. Pre-processed images that have been corrected for geometric distortions, bias fields and geometric scaling are available from the ADNI website, and were used in this work. Subjects were chosen with at least 2 scans with a minimum interval of 1 year. This minimum scan interval was chosen to minimise the possibility of errors in the estimation of TBM features as a consequence of the asymmetric registration model [140][203].

7.3 Experiments

7.3.1 Pre-Processing

Tools from the publicly available open-source software library, FSL[2] [171]were used to pre-process the images. Initially, all of the images were brain extracted using BET [170]. In many of the images, there is a large amount of neck voxels in the field of view. Therefore, BET was run with option "-B for bias field and neck clean-up" and "-f 0" to produce larger brain estimates. This succeeded in removing the majority of non-brain tissue, especially the neck, which tends to cause the most significant difficulty for the affine registration. To correct for differences in size and location, each of the baseline scans was registered to the MNI 152 template using FLIRT [98] with 9 degrees of freedom, and re-sampled to have 1mm isotropic voxels.

The follow-up images were brain extracted in the same manner as the baseline images. These were then rigidly registered, using 6 degrees of freedom to the baseline scan. The resulting transformation was composed with the transformation from baseline to atlas space, allowing a single interpolation from follow-up image to atlas space with 1mm isotropic voxels. The advantage of rigidly registering the longitudinal scan to the baseline, rather than directly to the atlas, is that regardless of any mis-registration to the atlas space, the images of each subject should be well aligned to one another. This will help ensure an accurate estimate of longitudinal TBM features.

Finally, in order to remove as much of the remaining non-brain tissue as possible, the MNI 152 template is registered to the baseline subject image with 12 degrees of freedom, to allow a more accurate affine registration. A brain mask, which is dilated by 2mm to avoid removing any brain tissue, is then propagated from the template space, to the subject space. This provides a sufficiently accurate, and consistent brain extraction, so that the non-brain voxels do not cause problems for the non-rigid registration.

7.3.2 Longitudinal Registration

The follow-up image is registered to the baseline image using the probabilistic registration algorithm with a global noise model. The hierarchical registration scheme used for the longitudinal registration is almost the same as that for the

[2]http://www.fmrib.ox.ac.uk/fsl/

previous experiments (given in Table 3.1). The only difference is the registration starts from level 2A. This is because there are no very large scale deformations, and starting from level 1A, where no deformations are required, can lead to λ being taken far from the optimal value. This is a problem because λ and ϕ for each level of the hierarchical registration scheme are initialised from the previous level. As VB provides an iterative local optimisation, if the initialisation is very poor, the optimal value may be too far away in parameter space to be found.

Registration Hyper-Parameters

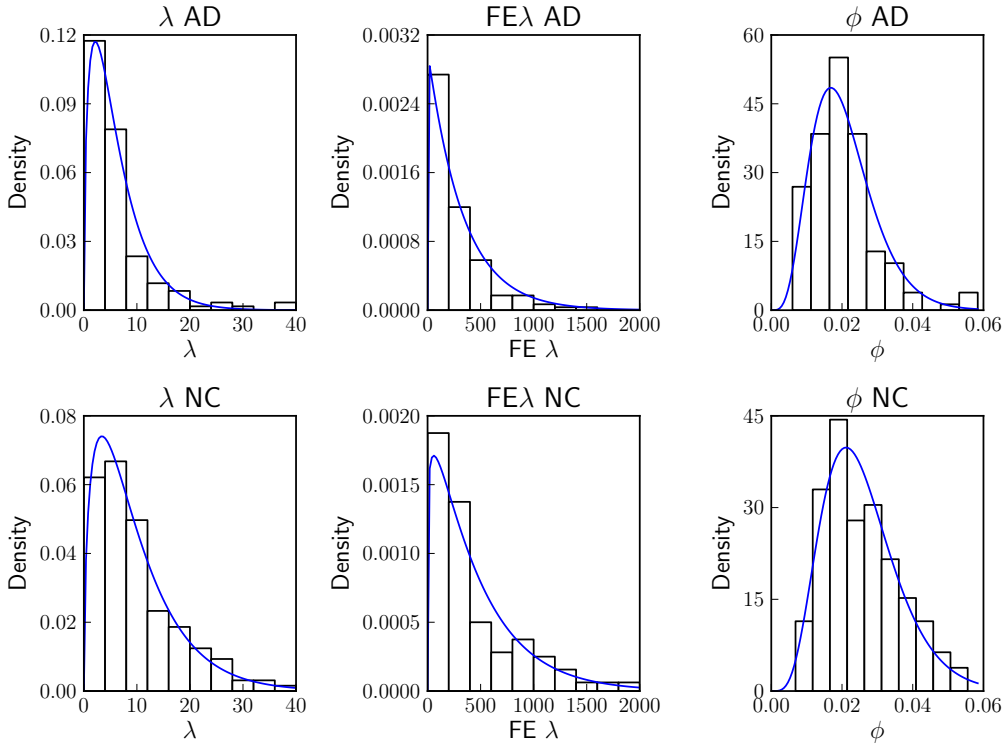

Figure 7.1: Registration hyper-parameters inferred from the longitudinal registration of data. The top row shows the distributions inferred for the subjects with Alzheimer's disease, and the bottom row shows the same for the normal control subjects. λ and ϕ for each of the population groups strongly resembles the estimated Gamma distribution, and shows no significant differences according to a Kolmogorov-Smirnov test at the 5% significance level. However, there is a significant difference for both λ and ϕ between the two populations as measured by the Wilcoxon ranksum at the 5% significance level. Larger λ and ϕ are inferred in the NC population than in the AD. The FNIRT equivalent λ (FEλ) shows that a very different level of regularisation is used than for inter-subject registration.

The hyper-parameters of the registration model at the final level of the hierarchical registration scheme are illustrated in Figure 7.1. The level of regularisation as given by FEλ, which was defined in section 5.5.2, is very different to that used in the FNIRT configuration for inter-subject registration. This is because longitudinal registration is a much simpler problem than inter-subject registration, and much smaller warps are required. This is why the spatial precision λ is so much higher, because the final transformation is much closer to the spatial prior. There are several subject registrations with very high λ values, these are from subjects where almost no differences are visible between the two images, and therefore the inferred transformation is very similar to the spatial prior. The values of ϕ are generally higher than those inferred in inter-subject registration, this is because there is a much better model fit.

Interestingly, the distributions of λ and ϕ differ significantly between the populations. A higher λ is found for NC, this is because there are smaller changes over time than in AD subjects, and hence the transformation is closer to the spatial prior. The distribution of ϕ is also higher for subjects with NC. This may be related to either more small scale changes in AD patients, which cannot be resolved without a more flexible transformation, or possibly the AD subject data containing more artefacts due to subject motion.

Longitudinal TBM Features

Analysis of the deformation field that maps between the two images allows the derivation of the determinant of the deformation Jacobian matrix. This provides a map of expansion or contraction of each voxel, and thus yields maps of estimated atrophy. The Jacobian determinants were constrained to be positive using the method of [102]. The logarithm of the determinant of the Jacobian is used in the experiments of this chapter. This is because it is symmetric, unlike the determinant itself [111]. The use of the logarithm means that expansion, or contraction by a given proportion, e.g. halving/doubling is equidistant from no volume change. This symmetry is important for statistical analysis to prevent a bias towards contraction, or expansion effects. The distribution of the logarithm should therefore be tighter, and less suceptible to skew resulting in improved statistical analysis across a population.

As the interval between scans varied across subjects, the logarithm of the determinant of the Jacobian values were linearly scaled to a single year.

An example of the benefits of using the probabilistic registration algorithm is given in Figure 7.2. Here, a longitudinal image pair are registered using the

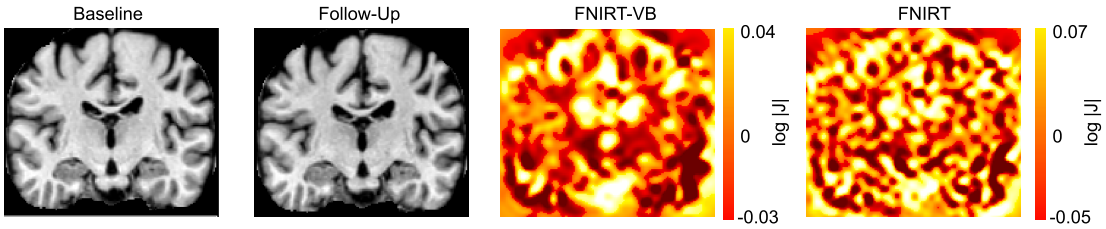

Figure 7.2: A longitudinal registration example of a normal control subject illustrating the difference in the TBM features between using inferred regularisation, and FNIRT using the standard scheme defined for inter-subject registration. The two structural MR images on the left are the baseline, and follow-up scans. The two maps on the right are of the log determinant of the transformation Jacobian matrix. The FNIRT-VB TBM maps appear very clean, whereas the FNIRT result is full of noise and artefacts as the registration is under-regularised.

proposed registration algorithm, or FNIRT with the standard FNIRT scheme for inter-subject registration. This is not a fair test, as no effort was made to find an appropriate parameter set for FNIRT for longitudinal registration. Nonetheless, it does further highlight a benefit of an adaptive approach to registration, as minimal parameter selection (just the starting level of the hierarchical registration scheme) was required.

Due to the unknown ground truth of deformation fields, no quantitative evaluation of the accuracy of longitudinal registration itself is presented. However, the discriminative ability of the Jacobian features derived from these deformation fields is evaluated in the following section. This can be considered as a surrogate measure of the accuracy of longitudinal registration, as the ability to discriminate between subject groups using the features is likely to be related to the biological accuracy of the registration.

7.3.3 Atlas Creation

The iterative framework proposed by Guimond et al. [75] as described in section 2.3.3 was used to create an atlas. The atlas is derived from 20 AD and 20 NC subjects in the training set. It should therefore be similarly representative of both populations, without inducing a bias in registration accuracy to a particular subject group. The atlas created using the probabilistic non-rigid registration algorithm is shown in several figures, but most clearly in Figure 7.4. The hierarchical registration scheme in Table 3.1 is modified for the atlas creation, and spatial normalisation experiments. The differences lie in a reduced level of image smoothing for the atlas image, as the atlas is already smooth. The modified

scheme smooths the atlas image using a Gaussian kernel with half the FWHM of the original scheme.

7.3.4 Spatial Normalisation

Each of the baseline images was registered to the atlas image to provide an accurate spatial normalisation for the TBM features.

Registration Hyper-Parameters

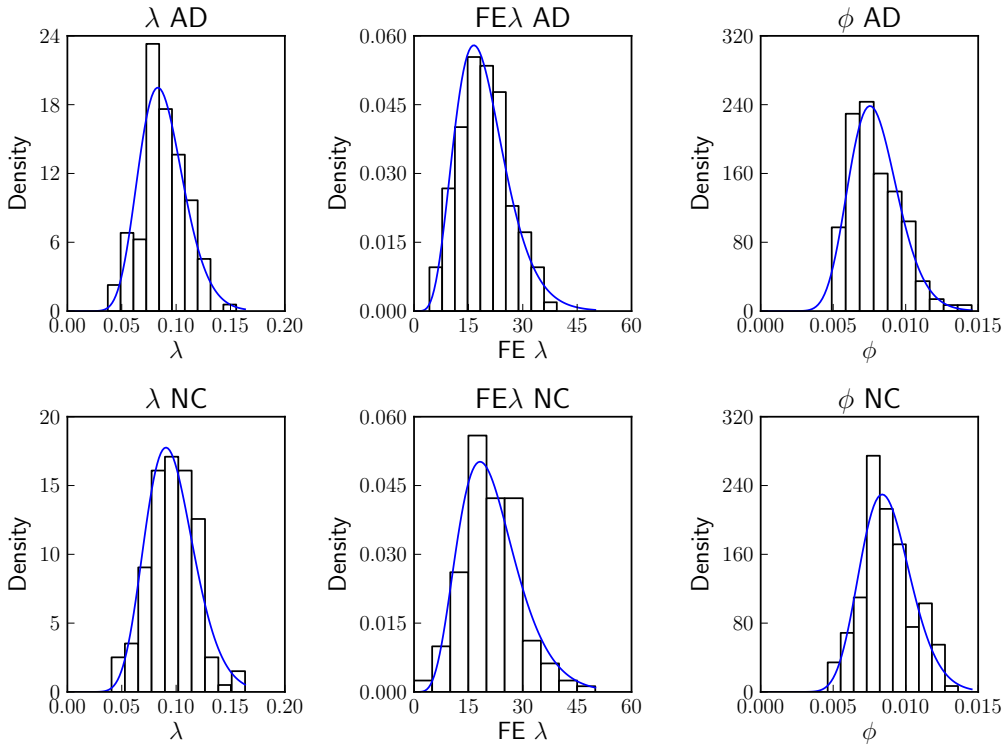

Figure 7.3: Histogram of the registration hyper-parameters inferred from the spatial normalisation of the baseline images of each subject. The top row shows the distributions inferred for the subjects with Alzheimer's disease, and the bottom row shows for the normal control subjects. Both the distributions of λ and ϕ for each of the population groups strongly resembles the estimated Gamma distribution, and shows no significant differences according to a Kolmogorov-Smirnov test at the 5% significance level. The distributions for each population have no statistically significant differences, as measured by the Wilcoxon ranksum at the 5% significance level.

The registration hyper-parameters inferred in the spatial normalisation of the

ADNI subjects are shown in Figure 7.3. The FNIRT equivalent λ values show a similar range to that inferred for the registration of the IBSR dataset. The λ values are slightly larger than those found for the IBSR dataset. This is because smaller deformations are expected as the images are registered to a representative atlas, rather than different subjects.

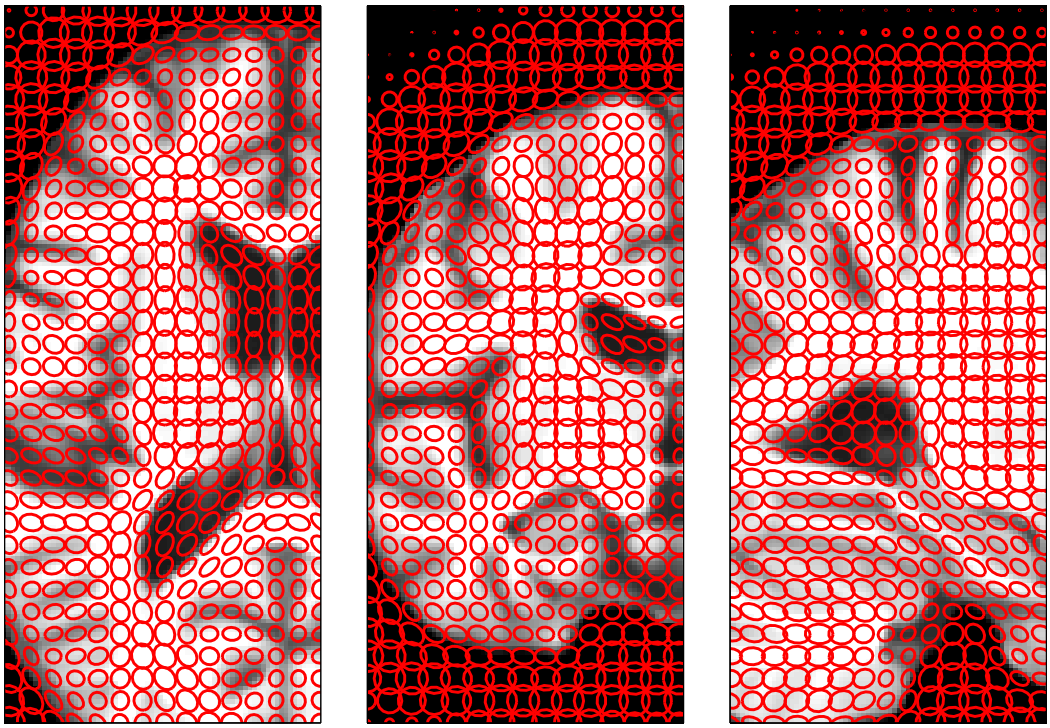

Figure 7.4: This plot shows the average voxelwise uncertainty distribution across all the spatial normalisations of the ADNI data to the atlas space. The uncertainty plot is overlaid on the atlas image. Each ellipse represents the full-width at half-maximum of the average uncertainty distribution in that plane at that location. The ellipse centres are plotted 5mm apart.

Registration Uncertainty

The probabilistic non-rigid registration algorithm was used to provide accurate spatial normalisation, which allows measurement of spatial uncertainty. The subject groups have a statistically similar distribution of λ and ϕ, the uncertainty distributions should be similar between the population groups. Hence, single

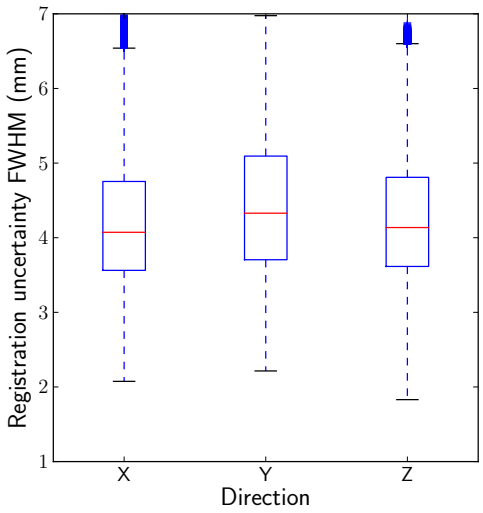

Figure 7.5: A boxplot of the average registration uncertainty across all the spatial normalisations of the ADNI data to the atlas space. Each column corresponds to the voxelwise distribution of the uncertainty in a given direction. As can be seen, the median of the uncertainty in all of the directions is quite close to a FWHM of 4mm, although there is a lot of variability across voxels.

plots using all the subjects are sufficiently representative. Figure 7.4 provides an illustration of the average registration uncertainty inferred by the probabilistic registration algorithm. This depiction allows the visualisation of the anisotropy of the registration uncertainty across the brain. To further explore and quantify the variability across voxels in the uncertainty distribution, Figure 7.5 shows a boxplot of the voxelwise average registration uncertainty in each direction. A map of the variability in the level of uncertainty across the set of spatial normalisations is given in Figure 7.6.

7.3.5 Spatially Normalised Feature Data

The longitudinal feature data can be accurately transformed to the atlas space using the inferred non-rigid transformation. This provides a set of feature data that can now be used for statistical analysis. Maps of the mean, and standard deviation of the spatially normalised longitudinal TBM data are given in Figure 7.7 and 7.8 respectively. As can be seen the difference in magnitude, and spatial location of the rate of anatomical changes between the two group averages is large. The differences are particularly strong in the temporal lobe region, ventricles and the cortex generally.

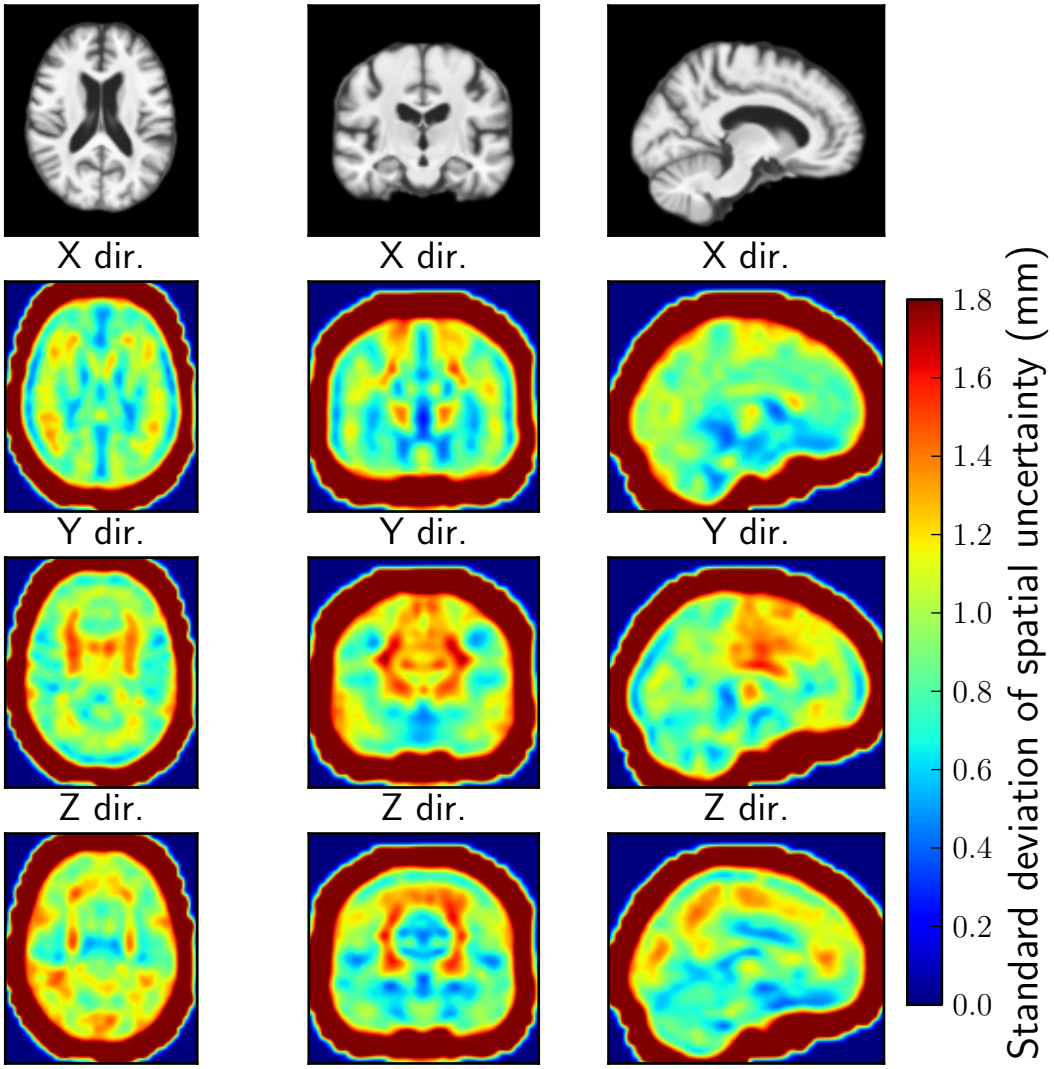

Figure 7.6: Plot of the standard deviation of registration uncertainty across the set of ADNI subjects that have been spatially normalised. The top row shows the atlas image, the subsequent rows show the standard deviation of the registration uncertainty in each direction. As with the IBSR registrations, the variability in uncertainty is highest in homogeneous regions.

The variability of the TBM features seems to be much greater in the AD population in general. For both populations, the region of greatest variability is in the CSF surrounding the cortex. This may be influenced through artefacts related to the brain extraction, which may or, may not keep the region of CSF at the edge of the brain. This region is expected to grow, as the brain shrinks, but it is unlikely to provide a robust feature due to this difficulty in pre-processing.

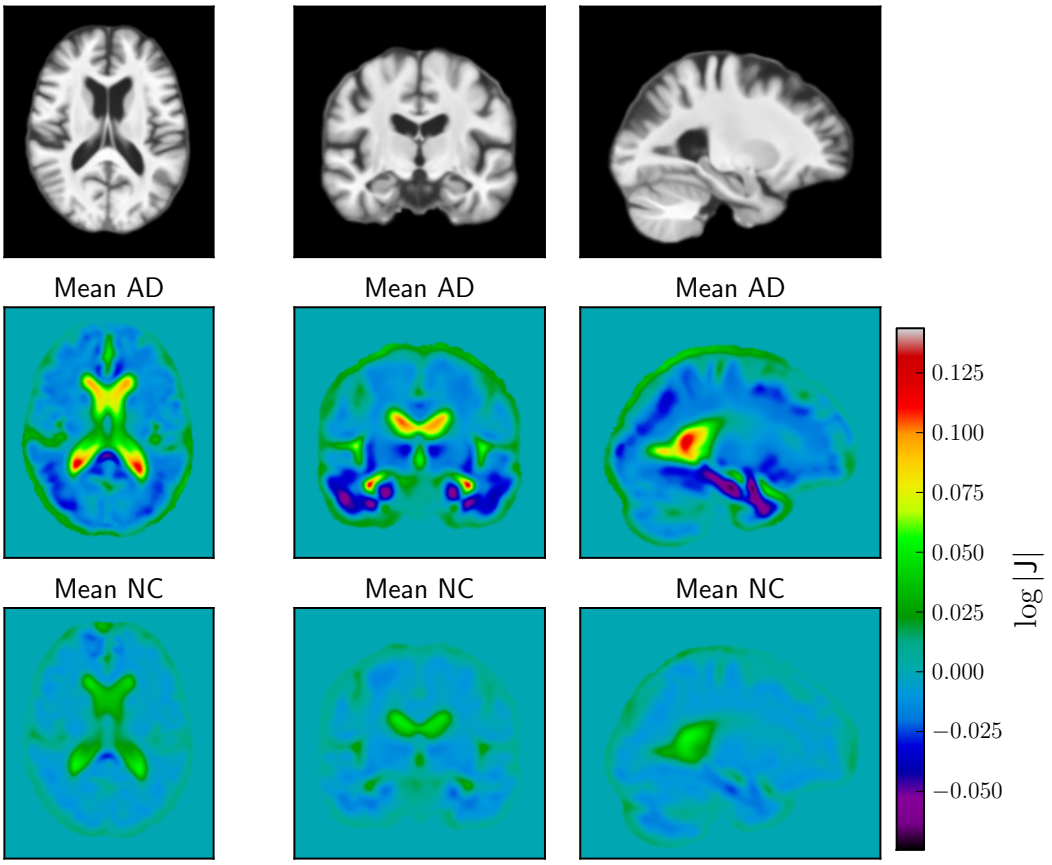

Figure 7.7: Spatial maps of the mean of the spatially normalised longitudinal TBM data across all the subjects in the training set. The top row shows the corresponding slices of the atlas image, the middle and the bottom row show the average TBM map for the AD and NC groups. There are visibly large differences in the rates of atrophy across the brain between subjects with AD and NC.

Interestingly, in the AD population the rate of expansion of the ventricles also shows a large variability. This may be because the rates of atrophy are non-linear across time [158], and the current state of disease progression will vary across subjects.

In the following experiments, the effects of smoothing this spatially normalised feature data are explored. Five different image smoothing levels are analysed, by either smoothing the data based on the registration uncertainty, not smoothing the data, or using a Gaussian kernel with a full-width at half-maximum (FWHM) of 2, 4, or 8mm. The different Gaussian kernels cover a range of expected signal sizes and levels of mis-registration.

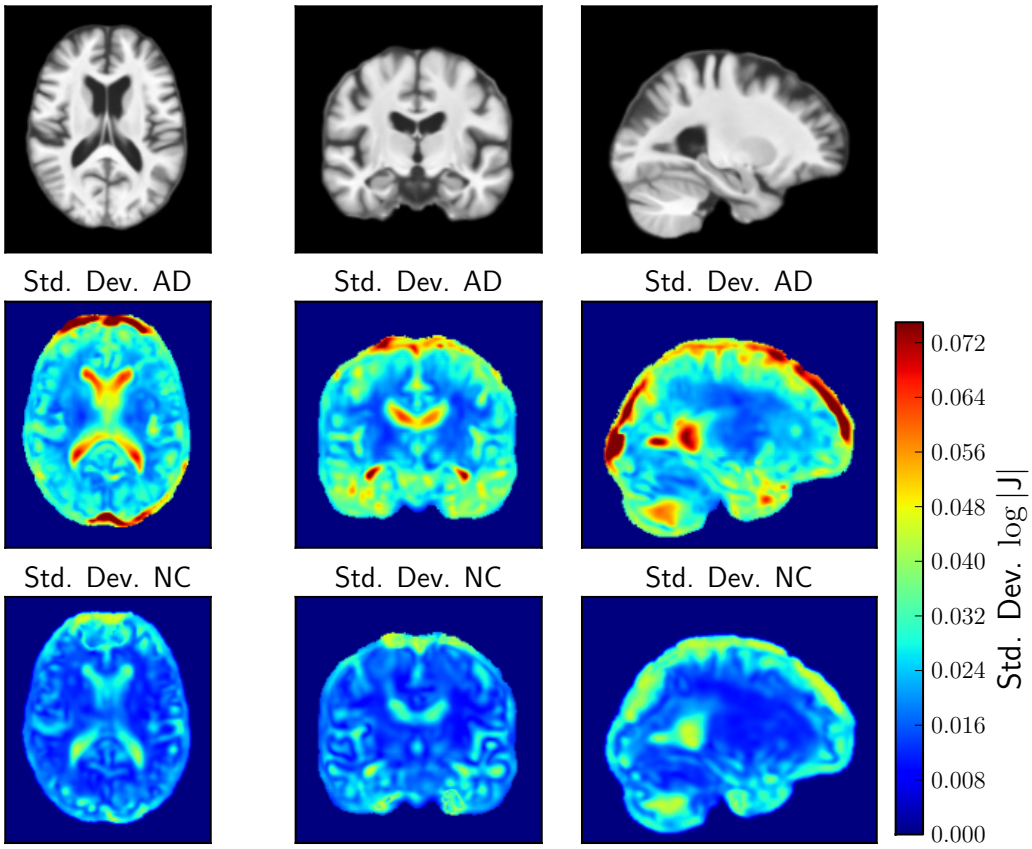

Figure 7.8: Spatial maps of the standard deviation of the spatially normalised longitudinal TBM data across all the subjects in the training set. The top row shows the corresponding slices of the atlas image, the middle and the bottom row show the map of standard deviation in the TBM data for the AD and NC groups.

7.3.6 Voxelwise Significance Tests

T-test Details

Voxelwise t-tests were run, using only the training set, to construct maps that illustrate the level of statistical significance in the difference of spatially normalised TBM maps between subject groups. This was performed for each of the levels of image smoothing. Two-sample two-sided t-tests are used to test the null hypothesis that data in a given voxel, for two groups, follows a distribution with equal means. The alternative hypothesis is that the means are different. T-tests assume that the data distribution for each population follows a Normal distribution, which for the case of the log $|\mathbf{J}|$ is approximately true [111]. However, as the variances of the two populations are different, as illustrated in Figure 7.8, Welch's t-test rather than Student's is used [190]. T-tests yield a test statistic that follow

a t-distribution if the null hypothesis is true. The statistical significance of a test statistic is measured with respect to the null hypothesis, as the probability, or P-value, that a test statistic of this value or less likely could have been randomly observed. A difficulty in applying t-tests in imaging data is that each voxel is treated separately. For 2 million voxels, as are in the atlas, extreme statistics are to be expected purely based on the null distribution. These are referred to as false positives, and they arise from the multiple comparisons problem. In this work, the multiple comparisons are corrected for using the false discovery rate (FDR), which is implemented as a tool within FSL. FDR controls the expected proportion of false positive voxels among the voxels that are deemed significant [131].

Results

Maps of the results of the t-tests under the different levels of smoothing are converted to z-statistics, and presented in Figure 7.9. All forms of smoothing increases the overall statistical significance of the data, as well as providing smoother estimates of significant regions. The use of the Gaussian kernel with 2 or 4mm FWHM and the uncertainty derived smoothing, all show patterns of the same shape as the un-smoothed data, but with increasing levels of statistical significance. Uncertainty derived smoothing, which has a median FWHM of approximately 4mm, reveals regions, including in the hippocampus, which are statistically more significant than the 4mm FWHM Gaussian. This may because there is less mixing of populations, due to the anisotropic nature of the smoothing in some regions, as was illustrated in Figure 7.4.

The use of the largest smoothing kernel (8mm FWHM Gaussian) leads to shape changes of the statistical significance map from the other approaches. For instance, the regions of CSF surrounding the hippocampi are shown to be less significant under large smoothing. This is because the expansion of these regions is over-smoothed with the hippocampal atrophy that is adjacent. Additionally, the shape of the region of statistical significance in the right temporal gyrus is greatly smoothed out. However, in these smoother regions the level of statistical significance is higher than with the alternative methods. This is highlighted in Table 7.2, which lists the most significant FDR corrected p-value, and its location, for each of the smoothing approaches. Interestingly, all methods except the 8mm FWHM Gaussian find the right hippocampus to be more statistically significant.

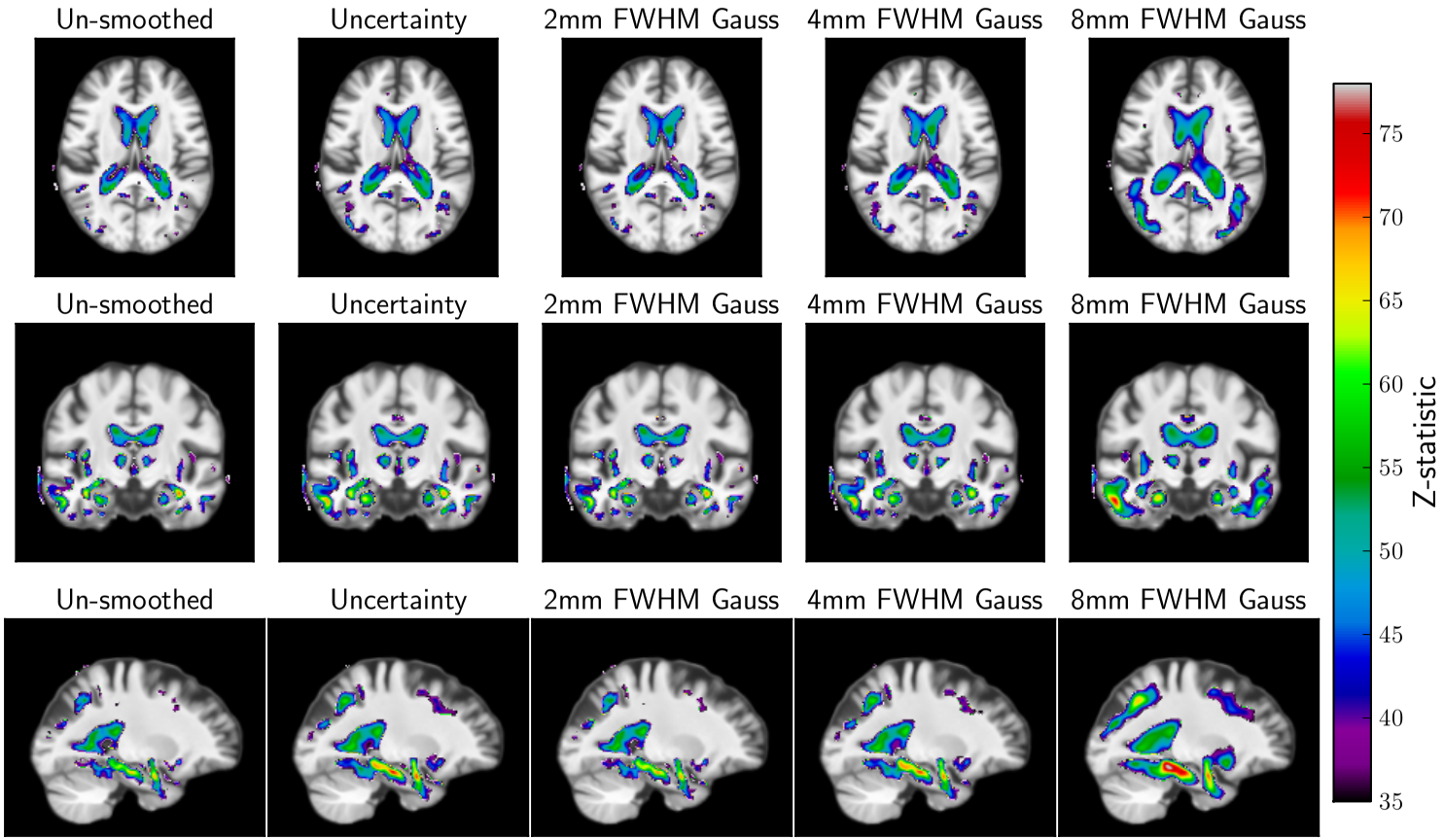

Figure 7.9: Thresholded Z-statistic map, derived from the corrected two sample t-tests comparing the two populations of training data. The thresholded z-statistic range is overlaid on the atlas image for the various levels of image pre-smoothing. Left and right are inverted for the coronal and axial slices according to convention.

Data	Location	FDR Corrected p-value
FNIRT-VB Un-smoothed	Right hippocampus	4.765×10^{-18}
FNIRT-VB Gaussian FWHM 2mm	Right hippocampus	3.304×10^{-18}
FNIRT-VB Gaussian FWHM 4mm	Right hippocampus	7.271×10^{-19}
FNIRT-VB Gaussian FWHM 8mm	Left hippocampus	1.986×10^{-19}
FNIRT-VB Uncertainty smoothed	Right hippocampus	3.835×10^{-19}

Table 7.2: Table of the location and significance level for the most statistically significant voxel between the two populations. The level of statistical significance is assessed through a two sample t-test between the two populations using the training data.

These results indicate that higher smoothing leads to voxels that are statistically more significant. Smoothing the data causes it to more closely resemble a Normal distribution, due to the central limit theorem. The central limit theorem states that the mean of a large number of independent and identically distributed random variables will tend to be Normally distributed. Thus, the smoothed data will more closely adhere to the assumption made by the t-tests. However, smoother data reduces the impact of smaller scale image features that may still be relevant to the prediction problem. It may also estimate artefactual stretching effects caused by the smoothing of two populations, which are similar in distribution but are anatomically distinct.

Previous studies into longitudinal TBM, such as Leow et al. [112] have focused purely on the voxelwise statistics to provide a comparison between alternative methods. However, voxelwise statistics do not tell the whole story. This is because voxels across the brain are often highly correlated. By analysing the data across multiple voxels simultaneously, more subtle voxelwise differences can be re-interpreted as distinct multivariate patterns. However, the computational cost of performing multivariate analysis may be greater than a voxelwise approach for certain methods. Additionally, the inclusion of irrelevant voxels may have a negative effect on the statistical analysis. For these reasons, feature selection may be reasonably applied. In this work, the t-test maps, which were calculated from the training set, were utilised as a means of selecting the most relevant voxels. The t-test maps maps were thresholded to select the 100,000 more statistically significant voxels from each level of smoothing. 100,000 voxels were chosen as it represents a sufficient number of voxels to encompass changes in several brain structures, as illustrated in Figure 7.10.

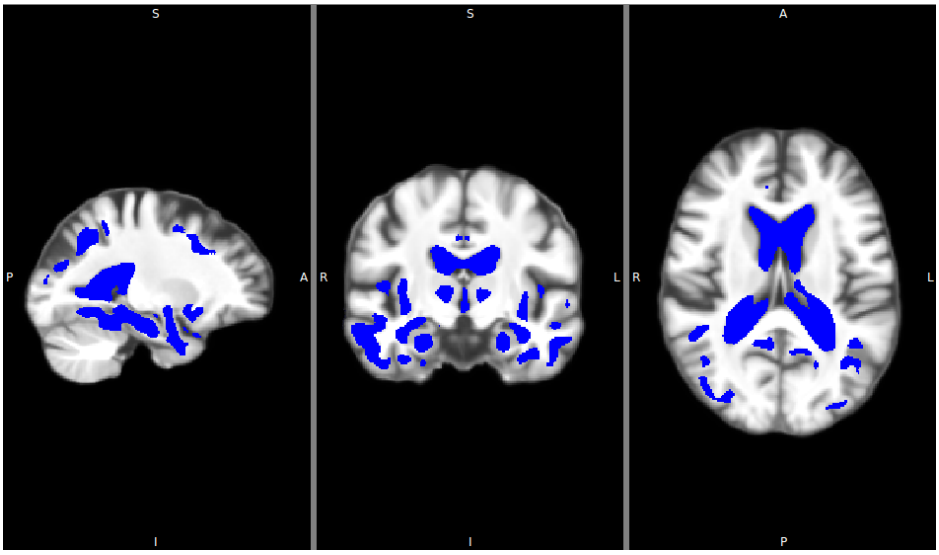

Figure 7.10: Voxel mask for the uncertainty smoothed TBM data in blue, overlaid on the atlas. The most significant 100,000 voxels were selected to be in the mask. Left and right are inverted for the coronal and axial slices according to convention. Voxels from the ventricles, hippocampi, temporal gyri, insular cortex and thalamus are all included in the mask.

7.4 Classification Experiments

To illustrate the discriminative ability of spatially normalised TBM features, as estimated by the probabilistic registration framework, and the effects of image smoothing, three classification methods are applied to the data.

7.4.1 Feature Preprocessing

Scaling and zero-centring of the data is commonly performed to enhance the ability to learn a true relationship between the data and labels. In this chapter, the data is scaled such that for each voxel, the variable values lie between -1 and 1, which is the approach recommended by the creators of LibSVM [87]. This procedure is performed to mitigate the effects of variable magnitude and variability, which might otherwise lead to a voxel being given undue significance by the classifier. It also increases the numerical robustness of the optimisation.

7.4.2 Classifiers

Two classification approaches are considered in this chapter, support vector machines, and Gaussian naïve Bayes, both of which were introduced in section

	Linear SVM	RBF SVM
Uncertainty Smoothed	$C = 10^{-2.5}$	$C = 10^{1.5}$ $\gamma = 10^{-5.0}$
Un-smoothed	$C = 10^{-3.0}$	$C = 10^{1.5}$ $\gamma = 10^{-5.5}$
2mm FWHM	$C = 10^{-3.0}$	$C = 10^{1.0}$ $\gamma = 10^{-5.0}$
4mm FWHM	$C = 10^{-3.0}$	$C = 10^{1.5}$ $\gamma = 10^{-5.0}$
8mm FWHM	$C = 10^{-4.0}$	$C = 10^{1.5}$ $\gamma = 10^{-5.5}$

Table 7.3: Optimal SVM parameters as inferred by 10-fold cross validation.

2.3.4. The use of linear and radial basis function kernels from the LibSVM package are experimented with. The optimal parameters for these models were obtained via 10-fold cross validation on the training set as described in section 2.3.4. The range of parameter values varied for each of the statistical models to cover the range of values that produce good results for the SVM models $C = [10^{-8}, 10^{-7.5}, ..., 10^{3.5}, 10^{4}]$ and the RBF kernel size $\gamma = [10^{-8}, 10^{-7.5}, ..., 10^{3.5}, 10^{4}]$. The optimal parameters for each dataset, for each of the SVM classification methods is given in table 7.3.

For the SVM classifiers, the choice of parameters can make a large difference to the performance of the classifier. Therefore the performance of the classifier on the testing data is dependent on the cross-validation. Whereas, the naïve Bayesian classifier has the advantage that no parameters need to be selected.

7.5 Classification Results

The classification sensitivity and specificity are illustrated in Figure 7.11. As can be seen, for all the classifiers the uncertainty smoothed data, and the 4mm FWHM Gaussian smoothed data obtain the same, and best level of classification. Interestingly, the 8mm FWHM Gaussian smoothing leads to a less accurate classification when using SVMs than any of the other methods. This is particularly interesting, as according to the voxelwise t-tests, the 8mm FWHM Gaussian would be expected to perform best. This may be related to the stretching of anatomically detected effects due to over-smoothing of the data. This is examined in Figure 7.12, where maps of the voxelwise SVM coefficient weights are plotted. It can be seen that the right temporal lobe, which has a strong classification weight in the uncertainty smoothed data, is much less important in the 8mm FWHM Gaussian smoothed data that places very high weight on a wide selection of voxels surrounding the hippocampus. Furthermore, heigher weights

are found around the edges of the ventricles, with lower weights in the middle. This is because only the boundary of the ventricles is reliably driven by the registration, and therefore finds consistent features, whereas the deformation of the interior comes from the smoothness imposed by the regularisation.

The optimal correct rate was found using the naïve Bayesian classifier, which gave a correct rate of 90.6%, and the naïve Bayesian classifier gave the best classification for all data types. A possible rationale for the improvement in classification when using naïve Bayes may be related to its explicit modelling of the different variances of TBM features between the two populations, and the lack of parameters that need to be tuned.

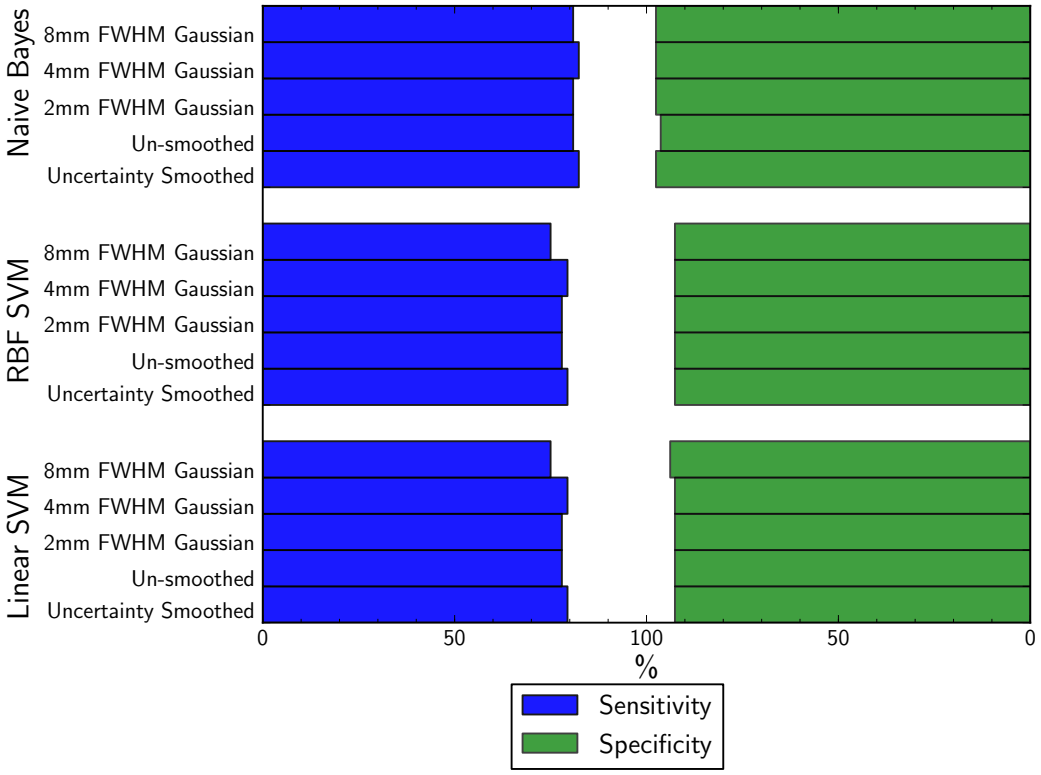

Figure 7.11: Stacked bar chart illustrating the sensitivity and specificity of the classification of Alzheimer's disease under different levels of data smoothing. Smoothing the data using the registration derived uncertainty, or the 4mm FWHM Gaussian achieves the most accurate classification for all methods. The naïve Bayesian classifier outperforms both SVM methods for all levels of smoothing.

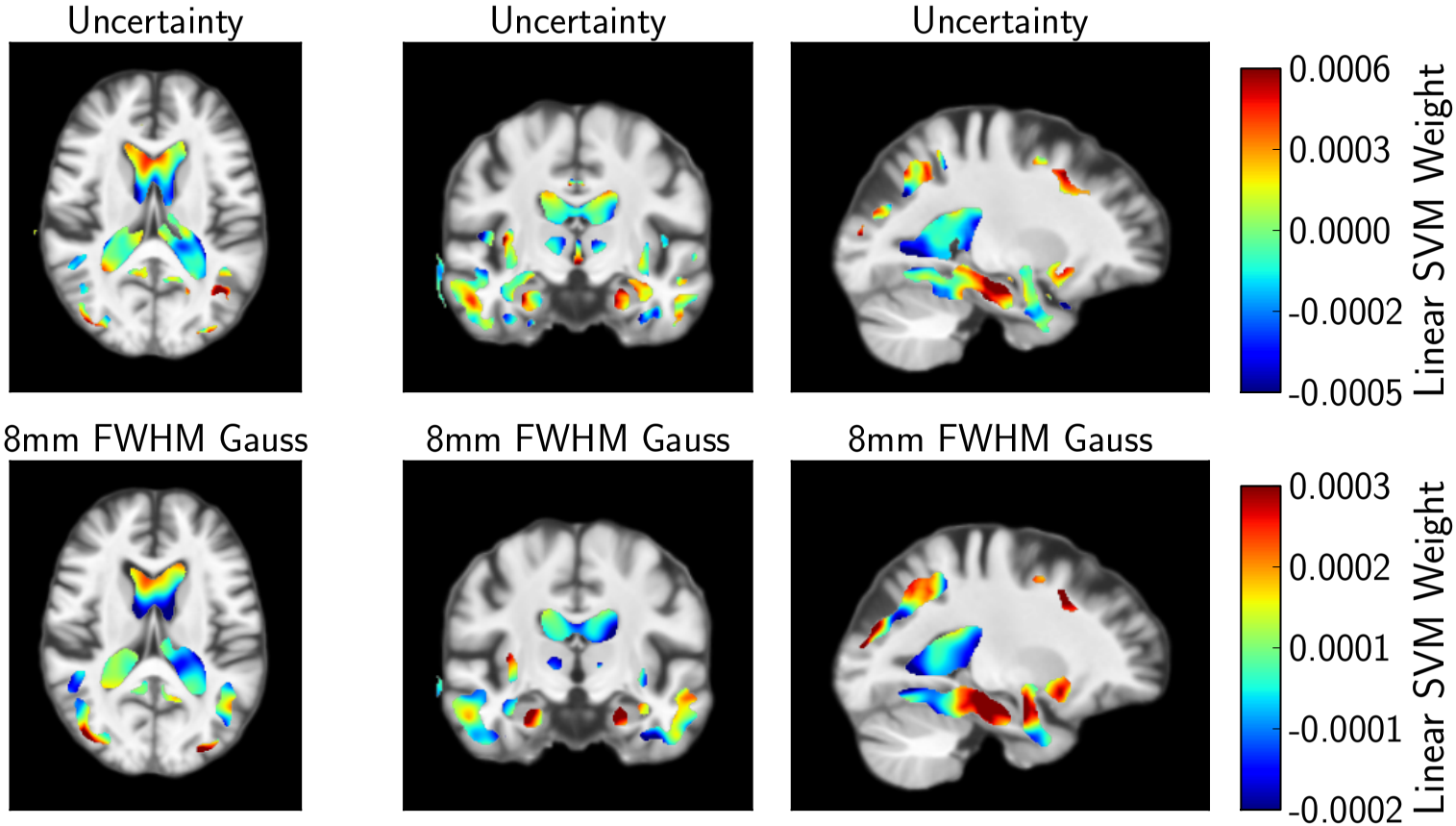

Figure 7.12: Linear SVM coefficients, where red indicates regions where contraction is indicative of AD, and blue showing regions which expand in AD as opposed to NC. Left and right are inverted for the coronal and axial slices according to convention.

7.6 Discussion and Conclusions

7.6.1 Discussion

The use of the probabilistic registration algorithm, with an inferred level of regularisation has been demonstrated to infer longitudinal TBM features that allow an accurate discrimination between subjects with AD and aged matched controls. Furthermore, using a smoothing kernel based on the registration uncertainty was shown to provide the highest rate of classification for all 3 classifiers. The use of a 4mm FWHM Gaussian achieved the same classification rate for all classifiers. This is contrary to what would be expected given the level of voxelwise statistical significance as assessed by t-tests. These found a higher level of statistical significance in the difference of the means between the two populations using an 8mm FWHM Gaussian. The uncertainty derived smoothing kernel, which provides on average a similar level of smoothing to the 4mm FWHM Gaussian, showed greater statistical significance than the Gaussian, but nevertheless gave the same overall classification performance. This implies that the differences between the two may not be relevant to this particular classification methodology.

Aside from registration uncertainty, the other reasons for image smoothing may influence the optimal smoothing method for data classification. The Normality of the data is influenced by the level of smoothing, and some classifiers (such as naïve Bayes), assume each voxel follows a Normal distribution for each population. Despite the fact that greater levels of smoothing should lead to the data being more normally distributed, the use of an 8mm Gaussian kernel produced lower classification accuracy using naïve Bayes than the uncertainty derived, or 4mm Gaussian smoothing filters. This may be due to over-smoothing of the data causing different populations to mix. The other rationale for image smoothing, based on the matched filter theorem, may also influence the optimal method of image smoothing. As demonstrated in Chapter 6, the use of the registration uncertainty derived smoothing filter leads to an improved correspondence, as measured by sensitivity and specificity of propagated subcortical segmentations. This may be due to the anisotropy of the filter, which is based on image information. If the shape of the signal of interest is related to the shape of the visible anatomical structures, than it could be expected than the registration derived uncertainty kernel may be closer to the optimal filter than an isotropic Gaussian.

An interesting extension to this work would be to fuse the TBM features from the subject to atlas registration, with the longitudinal TBM features. This could

allow a better estimate of disease status as both AD and normal ageing, are known to follow non-linear rates of atrophy [158]. Furthermore, this approach could be applied to subjects suffering from mild cognitive impairment, with the aim of evaluating whether longitudinal TBM features are predictive of conversion to AD. These are further discussed in chapter 9.

As TBM features are entirely derived from non-rigid registration, they suffer from any biases introduced by such algorithms. A particular case of this is highlighted in a discussion in volume 57 of Neuroimage in 2011. Hua et al. had originally described a method for analysing longitudinal estimates of atrophy in Alzheimer's disease from MR images [88]. Thompson and Holland [181] pointed out that this study showed biologically implausible deceleration of atrophy estimates from pairs of MR images with a short time interval (< 1 year). Hua et al. [91] acknowledged the presence of this bias, and recognised that it was due to the assymetry of the registration algorithm in use. Symmetric registration formulations have since been shown to resolve this problem [140][203], and this model could be extended to a symmetric formulation to ensure the estimation of robust TBM features. This is discussed in Chapter 9.

7.6.2 Conclusions

This chapter has shown that discriminative tensor based morphometry features can be derived from longitudinal data using the probabilistic non-rigid registration algorithm. The longitudinal features were spatially normalised to a representative atlas through the registration of the baseline image. The estimated registration uncertainty is demonstrated to provide an improved level of data smoothing, with respect to the task of classifying subjects. A maximum classification rate of 90.6% between subjects with AD and age matched normal controls was achieved using the naïve Bayes classifier, which outperformed linear or radial basis function SVMs.

The next chapter introduces an alternative approach to utilise registration uncertainty to improve the ability to predict a subjects disease status. Whereas this chapter investigated spatially normalised longitudinal morphometric features, the next chapter estimates differences between subjects based on a region of interest in a single image. Registration uncertainty is incorporated by drawing samples of probable registrations instead of data smoothing. Morphometric feature data from probable registrations are used to train several predictors that are combined in an ensemble learning framework.

Chapter 8

Ensemble Learning Incorporating Uncertain Registration

8.1 Introduction

This chapter describes a further approach to incorporating registration uncertainty into the prediction of Alzheimer's disease (AD) based on morphometric features. The use of a probabilistic registration tool provides an estimated posterior distribution of mappings, between subject and atlas space. Drawing samples from this distribution allows the estimation of a distribution of spatially normalised feature data, e.g. VBM or TBM features. Samples of spatially normalised feature data can be used as alternative training examples, which allows the creation of multiple statistical classifiers that can be subsequently combined using an ensemble learning approach. Furthermore, extra testing samples can be generated to measure the uncertainty of prediction. This is applied to classifying subjects with AD from normal controls (NC) based on a region of interest in structural brain MR images. This framework is applied using VBM and TBM features, and is compared to bootstrap aggregating, a common ensemble learning framework, in terms of classification accuracy, and the quality of the soft-classification predictions that arise from the ensemble. This work has been previously presented [166], and has been accepted as a journal paper [168].

8.1.1 Motivation

Medical imaging data is often used to make quantitative predictions about the current or future disease state of a subject. Morphological brain differences can be derived from structural MR images of the brain using VBM, or TBM as described in section 2.3. These features can be analysed across subjects to identify biomarkers, and construct models, which discriminate subjects with AD from NC. Machine learning techniques such as statistical classifiers and regressors are often used to facilitate this objective. These approach predict the value of an outcome variable, such as disease score, or group, on the basis of spatially normalised feature data.

The majority of machine learning techniques that are widely used for making predictions from medical imaging data are *supervised*, meaning that they require a set of training data with known outcome variables. This training set is used to derive a predictive model for estimating the mapping between feature data and outcome.

In any supervised learning approach, predictions on test data are highly dependent on the training data. As morphometric feature data are required to be spatially normalised prior to analysis, each item of training data is dependent on the registration between the subject and atlas space. As has been shown in the previous chapters, medical image registration is an intrinsically uncertain process, and a perfect spatial normalisation is implausible. Consequently, it is unlikely that any estimated statistical relationship derived from a given set of training data will be exactly correct, and any residual mis-registration of data may contribute to errors in prediction.

8.1.2 Previous Work

Several automated approaches have been demonstrated that allow the diagnosis and prediction of AD from VBM/TBM as described in section 2.3. However, minimal work has been published on the exploitation of estimates of registration uncertainty in statistical prediction. Very recently, Iglesias et al. [94] introduced an approach to hippocampal subfield segmentation, where registration uncertainty is integrated out of a combined registration/Gaussian mixture model approach to segmentation using MCMC. Risholm et al. [144] proposed an approach to calculating the uncertainty of a delivered dose in radiotherapy under uncertain registration. This work follows in a similar fashion by estimating the variability in statistical predictors that is due to the uncertainty in registration. The novel

contribution of this work is to leverage this variability within a statistical learning framework to provide a more robust prediction.

8.1.3 Proposed Approach

This work proposes to incorporate estimates of registration uncertainty within an ensemble learning approach [49]. Ensemble learning methods have been demonstrated to be an effective mechanism to measure the uncertainty of the space of statistical predictors, providing a more robust prediction by combining estimates. To provide variability in predictive models, they need to be trained using different training data sets. Each set needs to be selected appropriately to encapsulate a plausible level of variability in predictive models. In many settings, bootstrap aggregating or *bagging*, has proved itself to be an effective tool [29]. Bagging creates variability between predictors by sampling with replacement from the set of training subjects. However, the random selection of subjects in each training set can lead to large differences in predictors due to the omission of subjects. Conversely, this work seeks to leverage knowledge of the derivation of the data to create training data sets that encapsulate the intra-subject variability due to registration uncertainty.

In the case of spatially normalised feature data, the distribution of the data can be estimated from the distribution of probable mappings inferred by the registration algorithm. Samples can be drawn from the estimated feature data distribution for each training subject, and used as a parametric variant of bootstrapping [52]. These samples of feature data can be used in place of the MAP observations to build up a set of training data sets, which may contain all the subjects, but with examples based on different probable registrations. Such an ensemble of statistical predictors accounts for the inherent uncertainty in the registration process, and therefore leads to a more robust prediction.

This chapter describes how the distribution of feature data can be derived using a probabilistic registration tool, and how this can be incorporated into an ensemble learning scheme. This method is demonstrated for discriminating between subjects with AD and NC using an ensemble of linear SVMs, and for a combination of ensembles using data from different regions of interest.

8.2 Ensemble Learning Incorporating Uncertain Registration

8.2.1 Statistical Prediction

This chapter uses the notation that was previously introduced in section 2.3.4 to describe the statistical classifiers. As a brief recap: $\mathbf{d}$ is a set of feature data, e.g. TBM maps. o is an outcome variable, e.g. disease score. h is a statistical predictor, and $\hat{o}$ is an estimated outcome variable. A labelled training set of N data items is described as: $\mathcal{L} = \{(o_n, \mathbf{d}_n), n = 1, 2, ..., N\}$ where n indexes the subject. From $\mathcal{L}$, the predictors estimate the relationship between $\mathbf{d}$ and o. The trained predictive model provides an estimate for a test subject i based on $\mathcal{L}$, $\hat{o}_i = h(\mathbf{d}_i, \mathcal{L})$.

The relationship between a new test image $\mathbf{d}_i$, and its predicted outcome variable $\hat{o}_i$, is highly dependent on $\mathcal{L}$. Each training item of spatially normalised feature data $\mathbf{d}_n$ in $\mathcal{L}$ is dependent on the inferred image registration. Therefore, training a predictor using $\mathcal{L}$ is susceptible to mis-registration. Accordingly, mis-registration may contribute to errors in prediction.

8.2.2 Ensemble Learning

The novel aspect of this work is to incorporate the estimated registration uncertainty into statistical prediction using an ensemble learning approach. Ensemble learning methods [49] create a set of predictors to provide a more robust prediction. To explore the space of predictive models, they need to be trained using different data sets, $\{\mathcal{L}_m\} \subseteq \{\mathcal{L}\}$, each of length N, where m indexes the different training sets.

The class of ensemble learning methods that are considered use a linear combination of multiple statistical predictors to provide a more robust estimate:

$$\hat{o}_i = \Sigma_m^M \beta_m h(\mathbf{d}_i, \mathcal{L}_m) \tag{8.1}$$

where i is the index of a test subject, M is the number of predictive models, and β is a vector containing the relative weights attributed to each trained prediction model. Only binary statistical predictors are used in this work, although any form of predictor can be used. In this chapter, $\beta_m = \frac{1}{M}$, but alternative weighting

schemes could be used, including Bayesian model averaging [85]. The exploitation of the instability of predictive models under different training data should lead to an improvement in prediction for all but the most stable of predictors and data.

8.2.3 Bootstrap Aggregating

A standard approach to generating multiple predictors is bootstrap aggregating, or bagging [29]. In bagging, each $\mathcal{L}_m$ is selected by random uniform sampling with replacement from the set of training subjects. For a large number of bootstraps, each $\mathcal{L}_m$ would be expected to contain 63.2% of the unique training subjects [53]. This approach has been found to be effective at sampling the space of prediction models based on the inter-subject variability. A limitation with bagging is that it only considers the variation between subjects to create different predictive models. Whereas, the intrinsic uncertainty of the measurements from which the model is derived may lead to a comparably large and more reasonable source of variability.

8.2.4 Incorporating Registration Uncertainty

In this work, the knowledge regarding the derivation of the feature data is leveraged when creating $\{\mathcal{L}_m, m = 1, M\}$ such that it considers the distribution of feature data as estimated from the set of probable registration mappings, $P(\mathbf{d}|\Theta)$. This is achieved by selecting $(\mathcal{L}_m)_n$ to contain a random sample drawn from each subject's feature data distribution, $P(\mathbf{d}_n|\Theta_n)$. This is a parametric variant of bootstrapping [52], where instead of using observations, new data is drawn from the distribution of probable observations. This scheme is referred to as **Train+**. A graphical illustration of how multiple classification models can be generated in such a fashion is given in the top plot of Figure 8.1. The use of a sufficiently large ensemble of statistical predictors should account for the inherent uncertainty in the registration process.

$P(\mathbf{d}|\Theta)$ can also be used to provide additional information on the prediction variability for each predictive model. This is achieved by averaging the predicted outcome for a set of random samples drawn from the test subject distribution, $P(\mathbf{d}_i|\Theta_i)$, rather than simply testing using only the most likely observation. This scheme is referred to as **Test+**, which is graphically illustrated in the bottom plot of Fig. 8.1. **Train+** and **Test+** can be used separately, or can be combined together by using multiple test samples with each predictor in the ensemble. All

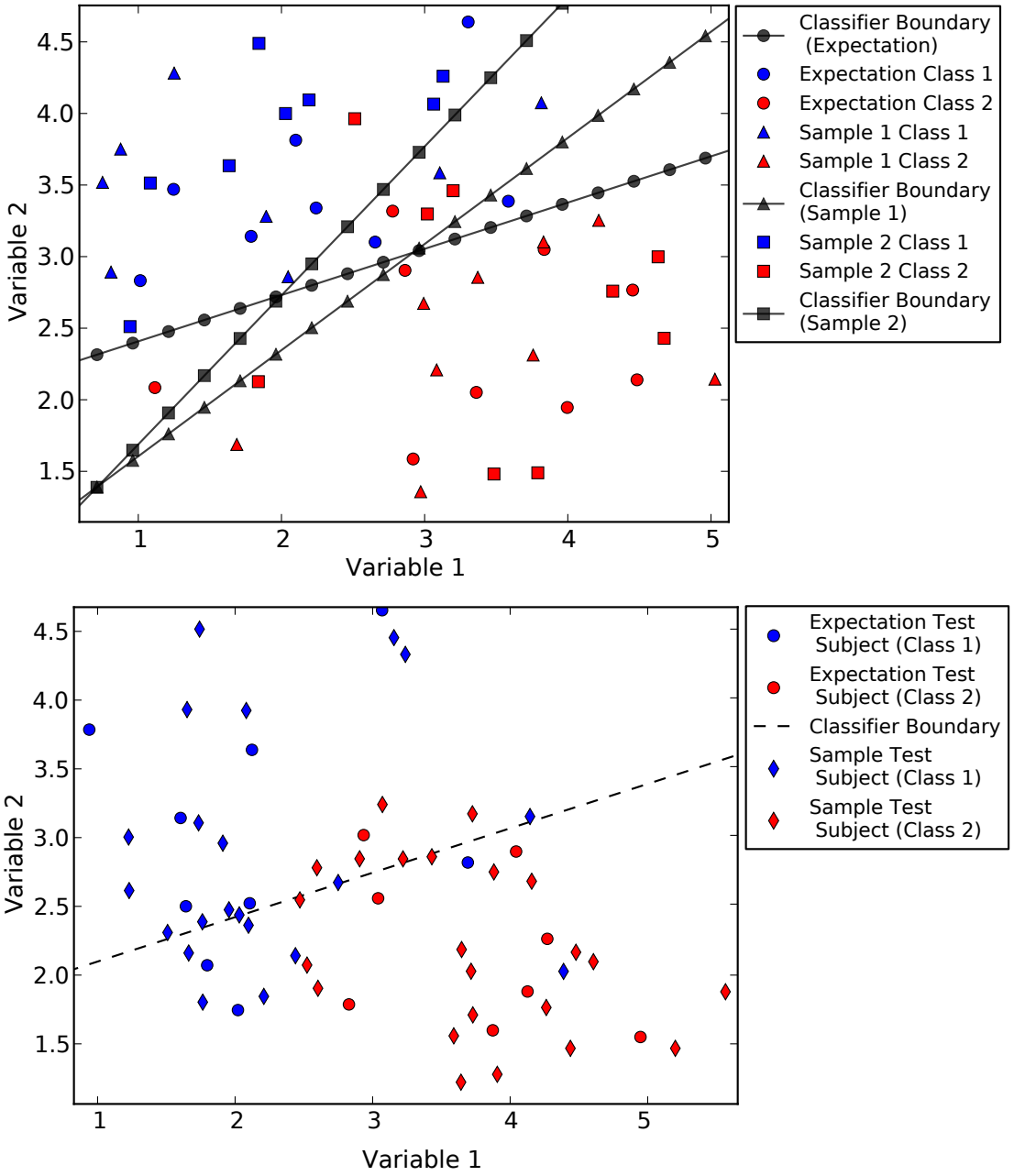

Figure 8.1: Graphical examples of how sampled data can be used in classification. The top plot illustrates the scheme **Train+**, where multiple classification boundaries are estimated using random samples of each subject. In this illustration two random samples are drawn from each subject, and thus three classification boundaries can be drawn, two from sets of samples and one using the distribution expectation. The bottom plot shows the scheme **Test+**, where the variability in classification label can be calculated using random samples of each test subject with a fixed classification boundary. In this case, three random samples were drawn for each subject.

of these variants can also be incorporated into a bagging framework, which would encapsulate both the inter, and intra-subject variability.

8.2.5 Feature Data

As previously described in section 2.3, VBM or TBM features can be used to describe subject brain morphometry. In conventional VBM, the segmentation probability map is transformed using the expectation of the transformation distribution, $\mathbf{d} = \mathbf{t}(\mathbf{x}, \boldsymbol{\mu})$, and modulated by the determinant of the deformation field Jacobian matrix, $\mathbf{J}_m$, to compensate for the expansion/contraction of voxels [71].

In TBM, instead of examining the spatially normalised image information, the assumption is made that the discriminative differences between subjects are contained in the deformation field that maps each subject to the atlas space [90]. Most commonly: $\mathbf{d}_v = \log |\mathbf{J}_m|_v$, where v is a voxel index. $\mathbf{J}_m$ is constrained to be positive using the method of [102].

8.2.6 Estimating a Distribution of Feature Data

A distribution of feature data for each subject $P(\mathbf{d}|\Theta)$, can be estimated for either VBM or TBM data. This is achieved by drawing samples from the inferred approximate posterior distribution of transformation parameters, $q(\mathbf{w})$, and calculating the resulting feature data for that sampled mapping. By sampling a large number of mappings, the distribution $P(\mathbf{d}|\Theta)$ can be numerically estimated.

8.3 Experiments

As before, data from the ADNI database was used in these experiments. The subject grouping is the same as in the previous chapter and is given in Table 7.1.

Pre-Processing

The data was skull stripped and affinely registered to the MNI 152 as described in section 7.3.1. Grey matter probability maps were estimated using FAST [205].

As drawing samples from multivariate normal distributions is computationally expensive for large number of transformation parameters, a region of interest (ROI) analysis was required. Atrophy in medial temporal lobe structures, particularly the hippocampus has been shown to be a sensitive marker of Alzheimer's

disease [97]. Therefore ROIs of 40x80x36 voxels surrounding the left and right hippocampi were extracted from the structural MR and grey matter images for registration.

Atlas Creation

A sharp atlas with minimal bias is estimated from all the subjects in the training set in an iterative manner as in [75]. This procedure is described in section 2.3.3. This is used to create a structural MR and grey matter atlas for each ROI.

Feature Generation

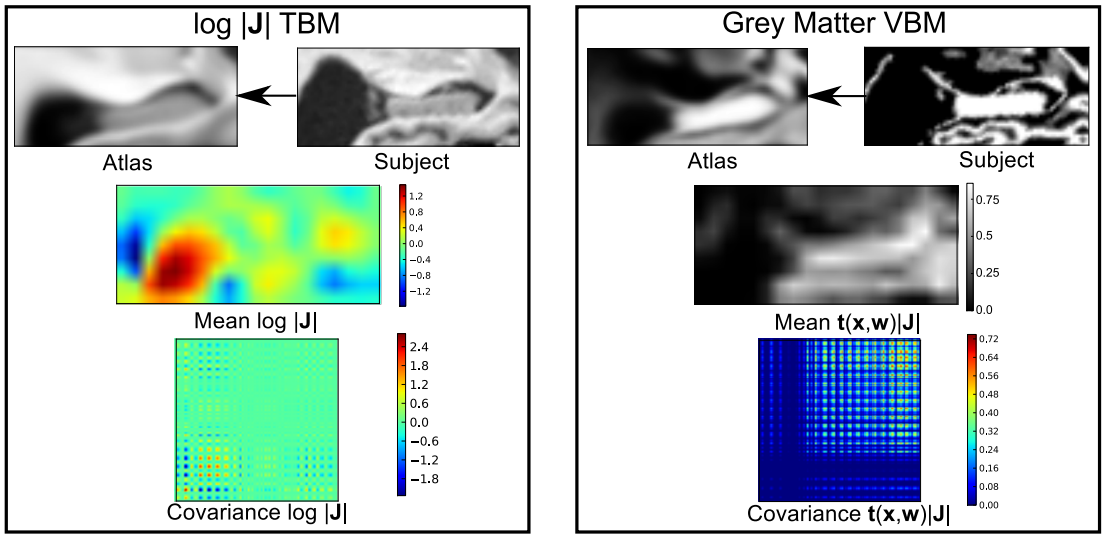

Figure 8.2: Examples of the data features acquired when registering a region of interest around the left hippocampus from a subject with AD taken from the test set, to the left atlas image for both TBM (left) and grey matter VBM (right). For both TBM and VBM a single slice of the volume is illustrated. The images marked Atlas and Subject are the high-resolution cropped Atlas and Subject images on which the registration is performed. The mean feature images show the mean of the estimated distribution for the same slice from the sub-sampled feature volume. The covariance matrices illustrate the estimated voxelwise covariance of the feature data for the displayed slice. The covariance matrix is ordered as $y * max(x)+x$ where x and y are positions in the feature image. The bottom left and the top right of the covariance matrix corresponds to the bottom left and top right of the feature image accordingly. The TBM data shows that there is expansion of the ventricle, but also that the log $|\mathbf{J}_m|$ in this region has a high degree of variance and covariance. The VBM feature data is a grey matter probability map that shows high variability across the image, except in the ventricles where there is no grey matter.

To create the feature data, each subject image was non-rigidly registered to the relevant atlas image using either a 5mm FFD knot spacing for TBM to give high-resolution features, or a 10mm spacing for VBM to align the images, but still retain subject differences. Once the registration algorithm has converged, warp samples are drawn from $q(\mathbf{w})$ to characterise $P(\mathbf{d}|\Theta)$. To avoid artefacts related to the edges of the ROI, the voxels within 4mm of the edge are removed, leaving a feature region of $32 \times 72 \times 28$ voxels. 4mm was chosen because the sampled deformations around the edge of the image are unlikely to exceed this.

To allow the tractable storage of samples from $P(\mathbf{d}|\Theta)$, the feature data needs to be sub-sampled by a factor of 4. This gives a total of 1008 voxels. To make the classification step computationally efficient, 3600 samples of the data feature are stored per subject to provide a sufficiently accurate description of the distribution, rather than fitting the data to a parametric distribution and later sampling from it. In the classification stages, samples are randomly chosen for each subject in both the Train+ and Test+ methodologies. In practice, all of these samples may be required for a run using Test+. An example illustration of VBM and TBM features for a subject in the test set with Alzheimer's disease is given in Fig. 8.2.

8.3.1 Classification

In the experiments a linear support vector machine (SVM) as implemented in LibSVM is used [33]. All 1008 feature voxels are used to classify between subject groups. The effects of subject age are regressed out on a voxelwise basis, this is to remove the effects of normal aging from the image features [51]. The regression is performed by fitting the general linear model:

$$\mathbf{y}_i = \mathbf{a}\beta_i + \epsilon \tag{8.2}$$

where $\mathbf{y}_i$ is a vector containing the de-meaned intensities at voxel i across the healthy controls in the training set. $\mathbf{a}$ is a vector containing the de-meaned ages of each subject. β_i is a scalar describing the linear effect of age on voxel intensity, and ϵ is an error term which we assume tends to 0. β_i can be calculated for every voxel using ordinary least squares fitting. Feature voxels for all subjects can now be corrected for normal aging by:

$$\mathbf{z}_i = \mathbf{y}_i - \beta_i\mathbf{a} \tag{8.3}$$

where $\mathbf{y}_i$ and $\mathbf{a}$ now contain both populations in the training and test populations, $\mathbf{z}_i$ is the corrected voxelwise intensity. Subsequently, each voxel in all subjects is re-scaled to take a value between -1 and +1 based on the statistics of the entire training data population.

Classifier Parameter Selection

Table 8.1: Linear SVM soft margin parameter C as selected by leave one out cross validation (LOOCV) using the expectation of the data features for each of the different data types.

Feature data	LOOCV correct rate	C
L $\log \lvert\mathbf{J}\rvert$	0.821	2×10^{-11}
R $\log \lvert\mathbf{J}\rvert$	0.821	2×10^{-12}
L VBM	0.877	2×10^{-10}
R VBM	0.827	2×10^{-10}

To select the most appropriate SVM classifier parameter for use in the Original scheme, a leave-one-out cross-validation procedure, as described in section 2.3.4, is used on the training set. This tests a range of the soft margin penalty parameter values, $C = [2 \times 10^{-15}, 2 \times 10^{-14}, ..., 2 \times 10^{15}]$, to find the value with the best generalisation accuracy. The value range in this procedure was selected in accordance with the recommendations of LibSVM[33]. The optimal parameter and its corresponding correct rate, defined as the ratio of correctly identified examples from the number of testing examples, are given in table 8.1. The optimal C parameter reflects the separability of the data classes. The found optimal values are all of a similar order of magnitude, at the small end of the scale, this indicates that the classifier can overfit on this data if too many support vectors are chosen.

For the SVM classifiers used within the ensemble learning schemes, $C = 2 \times 10^{15}$ that effectively removes the soft margin. This is beneficial as the training data is usually linearly separable, and removing the soft-margin introduces greater variability between classifiers as there is a greater dependence on the training data.

Ensemble Learning Schemes

In the classification experiments seven different ensemble learning schemes are compared:

Table 8.2: Classification correct rate using the different predictor training and testing variants using TBM and VBM feature data. L and R indicate the left and right hippocampus data, respectively. Naïve Bayes refers to the combination of soft probabilities from ensembles generated from different feature data types, and is described in Section 8.3.2. Train+Test+ and its variants tend to do as well, or better than a standard Bagging approach for both VBM and TBM. BaggingTrain+Test+ provides the best overall classification results.

Feature data	Original	Train+	Bagging Train+	Bagging	Test+	Train+ Test+	Bagging Train+Test+
L TBM	0.765	0.792	**0.799**	0.785	0.765	0.792	**0.799**
R TBM	0.718	0.7181	0.725	**0.738**	0.7181	0.7248	0.718
L VBM	0.826	0.846	0.852	0.852	0.826	**0.859**	0.852
R VBM	0.799	0.805	0.805	0.792	0.799	0.805	**0.812**
Naive Bayes All	0.718	0.846	0.852	0.839	0.832	0.846	**0.859**

- **Original**, train using the expectation of the whole training set and test using the expectation of the testing features.
- **Train+**, train using a random sample of each subject in the training set.
- **Bagging**, the standard bootstrap aggregating approach where the expectation of the examples in the training set are sampled with replacement.
- **BaggingTrain+**, where a random subject sample is used within a bagging scheme.
- **Test+**, train using the expectation of the whole training set and test using 20 random samples for each subject in the test set.
- **BaggingTest+**, the combination of bagging and test+.
- **Train+Test+**, the combination of train+ and test+.

In these experiments 300 classification models were generated to make an ensemble as this was sufficient for convergence for all methods. A summary of the results of these experiments is given in Fig. 8.3, and the classification correct rate for all of the methods is given in table 8.2.

As shown from table 8.2, the left hippocampus provides stronger features than the right for discriminating between Alzheimer's disease and age matched normal controls for all methods and both feature types. VBM provides good separation for both left and right hippocampi, whereas TBM of the left hippocampus provided substantially better discrimination than the right. Fig. 8.3 provides a

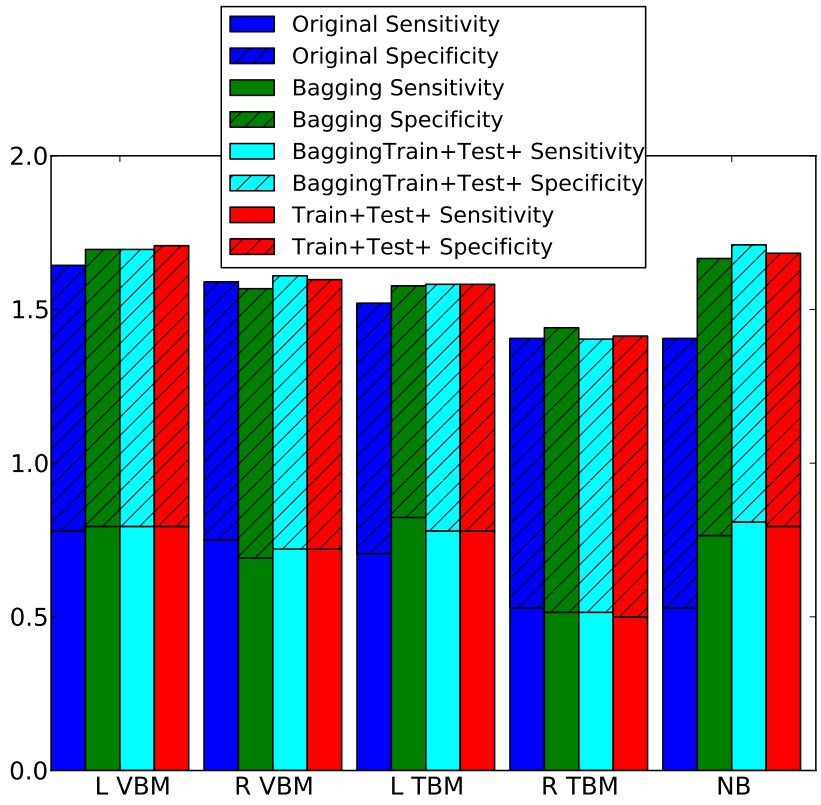

Figure 8.3: Stacked bar chart illustrating the sensitivity and specificity of the classification of Alzheimer's disease in selected experiments. L and R represent left and right ROI images respectively. NB refers to the results of the naive Bayes combination of all the data features in a post-classifier analysis as described in Section 8.3.2. It can be seen that BaggingTrain+Test+ and Train+Test+ outperform, or do as well as standard Bagging for all the features types, except R TBM, which is the lowest performing feature for all methods. BaggingTrain+Test+, when combining ensembles using naïve Bayes, gives the best trade-off of classification sensitivity, and specificity of any of the approaches considered.

summary of some of the results illustrating the sensitivity and specificity of each ensemble. Sensitivity is defined as the proportion of correctly identified disease cases, out of the total number of disease cases. Specificity is the proportion of correctly identified control subjects, out of the total number of controls. This summary shows that ensemble learning approaches generally provide more accurate classification than the Original approach. All of the Train+Test+ approaches outperform, or do as well as Bagging for all features except right TBM, which produced the lowest classification results for all methods. This implies that the additional data variability provided by registration uncertainty assists in creating a more accurate ensemble.

Furthermore, the largest improvement over the original approach is found in the left hippocampus TBM feature data, particularly using the BaggingTrain+ schemes. The strength of this improvement is likely to be due to the data feature being derived from the warp field. Small changes in warp field that might have little effect on the image likelihood, may lead to more substantial changes in $\log|\mathbf{J}_m|$. This is likely to also be a contributing factor in the improvement in the VBM results, as the warped GM probability map is modified by the $|\mathbf{J}_m|$. The Test+ schemes do not have much impact on classification correct rates when using a single ensemble, but help when combining ensembles using Naïve Bayes.

Computational Cost

In terms of computational time, the Original scheme is fastest, as only 1 classifier is constructed. Including the overhead of loading, and pre-processing the data, the training takes approximately 5 seconds. The use of 300 bootstrapping samples in Bagging takes 5 minutes to complete. Train+ and BaggingTrain+ take about 5 minutes of CPU time. However, because of the slow speed of disk access that is required to load the samples, Train+ and BaggingTrain+ take 10-15 minutes. The use of Test+ schemes adds an additional 30 minutes to run-time. This extra time is almost entirely taken up by disk access. Once the ensembles have been created, classification of a new sample is very fast, ≈ 1 second, and 10 seconds using Test+. All of the experiments were conducted on a dual core 2.8GHz laptop with a serial ATA (7200RPM) hard-drive.

8.3.2 Combining Soft Classification Probabilities

Soft Classification Probabilities

Each ensemble gives a soft classification result for each subject as it is an average of multiple predictions, given by:

$$P(o_i|\mathbf{d}_i^j, \{\mathcal{L}^j\}) = \Sigma_m^M \frac{1}{M} h(\mathbf{d}_i^j, \mathcal{L}_m^j) \tag{8.4}$$

where h is restricted to being a binary classifier and the superscript j denotes the feature data type. An illustration of the soft classification resulting from different ensembles in given in Fig. 8.4.

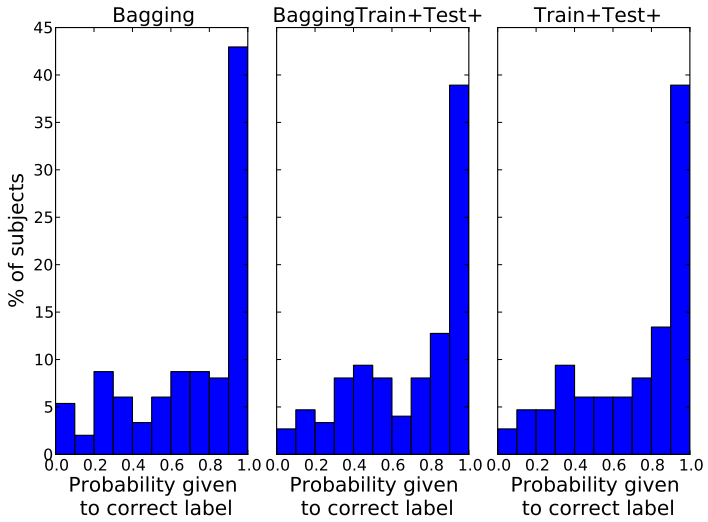

Figure 8.4: Histogram of the probability estimates given by a classification ensemble, to the correct class label. These histograms are plotted for the right hippocampus TBM feature data. As can be seen, the different methods of incorporating classifier variability induce different levels of classification uncertainty. In this example, Bagging assigns a 70% or more confidence to 16 subjects that it wrongly classifies, as opposed to 12 from BaggingTrain+Test+ and 13 in Train+Test+.

Naïve Bayes Combination

To assess how reasonable the soft predictions given by each classifier ensemble are, they can be combined in a post-classifier analysis. Naïve Bayes provides a simple framework for combining probabilities, assuming that they are independent, which is true for the soft classification probabilities presented here. Therefore, the soft classification probabilities are combined using a naive Bayesian classifier:

$$P(o_i|\mathbf{d}_i, \{\mathcal{L}\}) = \frac{\Pi_j^J P(o_i|\mathbf{d}_i^j, \{\mathcal{L}^j\})}{\Pi_j^J P(o_i|\mathbf{d}_i^j, \{\mathcal{L}^j\}) + \Pi_j^J P(\neg o_i|\mathbf{d}_i^j, \{\mathcal{L}^j\})} \tag{8.5}$$

where $\neg$ indicates the alternative label in a binary classification problem. Results for the naïve Bayes combination are provided in Table 8.2 and Fig. 8.3. The combination of Bagging with Train+Test+ leads to the most reasonable soft-predictions, as shown by the highest correct rate and best sensitivity/specificity trade-off. All of the Train+ schemes also outperform Bagging in the naïve Bayes classification. This is likely to be caused by Bagging showing slightly greater over-confidence in incorrect predictions, as illustrated in Fig. 8.4. It can be seen that the use of Test+, on its own, or in combination with BaggingTrain+ is

beneficial, implying the additional testing samples help estimate the uncertainty of prediction.

8.4 Discussion and Conclusions

8.4.1 Discussion

The incorporation of registration uncertainty into an ensemble learning scheme for statistical classification, has been demonstrated to generally improve classification compared to a standard scheme or bootstrap aggregating. These experiments were conducted using features from grey matter voxel based morphometry, and tensor based morphometry. These results imply registration uncertainty contains more useful information for the discriminative problem, than that obtained from bootstrapping. Furthermore, it was shown that the variability induced by bootstrap aggregating leads to ensembles with less reasonable estimates of prediction uncertainty than the proposed approach. This was demonstrated through the use of naive Bayesian combination of ensembles that are created from different data features.

An alternative approach to utilise the full estimated distribution of registration mappings would be to use the transformation parameters themselves as data features. As the inferred distribution is multivariate normal, there is a finite length description of the distribution. This description could be used directly, rather than through sampling the distribution. A disadvantage of using such an approach is that some of the interpretability associated with VBM/TBM is lost when using the transformation parameters directly as features.

Posterior probabilities can be directly estimated from a variety of classifiers, including SVMs, but in a more principled manner from logistic regression or relevance vector machines [182]. Such probabilistic classification estimates, or regression outputs, could be incorporated within the proposed framework.

Further work could be carried out to improve the overall classification accuracy to match current state of the art pipelines. Firstly, multiple functional areas could be considered, and their soft-classification probabilities combined. A more flexible ensemble learning scheme could be used, such as Bayesian model averaging. Feature selection could be used in place of, or following voxel sub-sampling for creating the feature data distribution, to select the most discriminative voxels. In particular, the choice of classifier, and its associated parameterisation could be addressed as well as the data processing scheme.

8.4.2 Conclusions

This chapter has describe a framework in which registration uncertainty can be incorporated into an ensemble learning scheme to provide more accurate prediction, with more reasonable estimates of classifier uncertainty, than standard approaches such as bootstrap aggregation. This was achieved by sampling probable registration transformations inferred from a probabilistic registration algorithm, and then estimating the distribution of a data feature given the uncertainty in the registration. Samples of the feature data distribution are used in place of the most likely observations in the training and testing phase for statistical predictors. The proposed approach generates prediction variability from the intra-subject uncertainty as opposed to the inter-subject variation, as is achieved by boostrapping. This chapter has described a method of combining predictors trained using sampled data into an ensemble. The experiments provide results on the problem of classification of subjects with Alzheimer's Disease, from age matched healthy controls using a linear SVM.

The final chapter provides some conclusions, and ideas for future related work.

Chapter 9

Conclusions and Outlook

9.1 Summary of Contributions

This thesis has presented a novel probabilistic model, and inference scheme for non-rigid registration of medical imaging data. This framework permits the data-driven inference of the level of regularisation, and provides estimates of the uncertainty in registration. This framework has been applied to detecting discriminative differences, and changes over time, in brain morphology between MR images of subjects with Alzheimer's disease (AD), and age matched controls. The major contributions of this thesis are outlined below.

Probabilistic Registration Framework with Inferred Regularisation

Chapter 3 described a novel probabilistic non-rigid registration model, and a tractable full Bayesian inference scheme. By modelling the regularisation, and data fidelity level within a hierarchical Bayesian model, these parameters can be inferred from the data alongside the transformation parameters. This allows the registration framework to be highly adaptable, and capable of being applied to a variety of data with minimal, or no parameter selection. This model was extended in chapter 4 to allow for a spatially variable description of noise. In chapter 5, these frameworks were evaluated for the problem of inter-subject registration, in terms of accuracy of anatomical label propagation, and transformation smoothness. The proposed registration frameworks were shown to provide accurate reg-

istration of a range of cortical, and sub-cortical anatomical structures. A wide range of transformation complexities were found due to the data-driven nature of the regularisation. The adaptability of the regularisation and data fidelity parameters was shown to be beneficial where differences in image acquisition made matching difficult.

This registration framework was also applied to the spatial normalisation of structural MR brain images of subjects with AD, and age matched healthy controls in chapters 7 and 8. Furthermore, in chapter 7 this framework was used to describe longitudinal morphological changes in subjects with AD and controls. The derived morphological features were demonstrated to accurately predict whether a subject suffers from AD. The flexibility of applying this algorithm to different problems helps highlight the benefits of an adaptive approach to registration.

A further advantage of the probabilistic registration framework, is that it intrinsically provides estimates of the uncertainty in the registration parameters; most notably, in the parameters that describe the transformation.

Registration Uncertainty Derived Smoothing

Chapter 6 discussed the estimates of registration uncertainty given by the Bayesian registration algorithms, and introduced an adaptive data smoothing filter that is derived from the registration uncertainty. Smoothing of spatially normalised data was motivated in terms of compensating for residual mis-registration in propagating anatomical segmentation labels. The smoothing filter derived from registration uncertainty, is demonstrated to better compensate for residual mis-registration than isotropic Gaussian smoothing kernels. In chapter 7, the registration uncertainty derived smoothing filter was shown to improve the predictive capability of spatially normalised longitudinal Jacobian features, achieving the highest rate of classification between subjects with AD and controls.

Ensemble Learning Scheme Incorporating Registration Uncertainty

Chapter 8 introduced a method for exploiting the estimated registration uncertainty for improving statistical prediction in an ensemble learning framework. Samples of morphological feature data are estimated by drawing probable registrations from the posterior distribution of transformation parameters, as estimated by the probabilistic registration framework. These samples are used to create an array of classifiers, which are each trained using different sampled

registrations. These classifiers are amalgamated together in an ensemble learning framework. Ensembles created using registration uncertainty are shown to improve the accuracy, and uncertainty of prediction compared to bootstrap aggregation, which is commonly used in ensemble learning.

9.2 Directions For Further Research

9.2.1 Registration Framework

The probabilistic registration model, and inference scheme developed in this thesis could be extended in several ways to provide more accurate and robust registration, and to potentially provide more accurate estimates of registration uncertainty.

Choice of Transformation Model and Formulation

The biggest limitation of the presented registration model, in terms of registration uncertainty, lies in the mean-field approximation in the inference of ϕ and $\mathbf{w}$, $q(\mathbf{w}, \phi) \sim q(\mathbf{w})q(\phi)$. As described in section 6.4.1, the approximation of independence means the shape of the image derived uncertainty is necessarily provided by a single image, as opposed to the joint distribution of $\mathbf{w}$ and ϕ, which would draw information from both images. Unfortunately, the calculation of the joint distribution itself would most likely be intractable. However, this situation could be largely improved through the application of a symmetric approach to registration. In such a methodology, two transformations could be inferred, one from each image, to an unbiased average space between images. Each of these transformations could still be described using a multivariate Normal. This would lead to a more complex uncertainty distribution of the mapping between the two images, which is dependent on both images.

Such an approach would be difficult to implement within a small deformation framework, as both images are registered to an estimated centre point in image space. Calculating the transformation from one image to another requires two transformations to be applied. One of which must be an inverse of an estimated mapping,. The inverse of a mapping inferred by a small deformation model is not guaranteed to be well defined. Furthermore, it will not necessarily be well described using a parametric form. Even greater complications arise from the notion of inverting the parametric uncertainty information.

A symmetric registration approach would be better considered through the use of a velocity field transformation model. In such a model, inversion is a trivial process of following the velocity field backwards. Likewise, the uncertainty of each velocity component may be more trivially invertible, especially if modelled using a multi-variate Normal, due to the symmetry of the distribution. However this would require further investigation to validate. Regardless, the implementation of this framework within a symmetric, large deformation mode, would allow a more interesting, and potentially more fruitful, analysis of registration uncertainty; although the increase in transformation parameters will come with a computational penalty.

Derivation of an Adaptive Spatial Prior

The use of a biologically appropriate spatial prior is likely to yield an improvement in both the registration results, and the nature of the registration uncertainty. This is because instead of penalising bending energy, it would penalise deviation from the distribution of probable brain registrations. This would aid the use of any methods utilising the estimated registration uncertainty, as any samples from the prior would be possible brain registrations, rather than simply smooth mappings. As described in section 2.1.3, there are a variety of approaches for creating a biological, or adaptive prior that could be considered. Although deriving a prior from the covariance matrix of many registration is an attractive solution, it is likely to be computationally infeasible due to the dense nature of the covariance matrix. The use of such a prior would also require the transformation parameters for each registration to be defined in a common space, which is a limitation. Although in principle, such an approach could be taken in groupwise registration as in [2]. The approaches of Friston et al. [62] or Harrison et al. [78] that allow the data-driven inference of a covariance structure, may be more reasonable to investigate, perhaps in conjunction with an approximation of independence between regions [79].

Alternative Noise Models

The local noise model, presented in chapter 4 is likely to perform better with a more flexible and less spatially smooth transformation model. However, the use of alternative noise models, which may be a better fit to the residual distribution could be experimented with. This could involve perhaps using a mixture of Gaussians to model the outliers and well matched voxels separately. Alternatively,

an approach similar to that of Zöllei et al. [208] could be used to model the joint image distribution directly, and the uncertainty estimates from different noise models could be compared.

Bayesian Model Comparison

Although computationally expensive for the registration framework presented in this thesis, Bayesian model comparison is an objective approach to compare models where an appropriate gold standard is not available. This could be used to provide an objective comparison of different model extensions, to determine whether the benefits of a more complex model, outweigh the additional complexity.

9.2.2 Statistical Analysis in Alzheimer's Disease

The most important consideration for future statistical analysis of AD, and its progression is the inclusion of subjects with mild cognitive impairment [136]. In particular, these subjects should be investigated for biomarkers that are predictive of conversion to AD.

Combining Atlas and Longitudinal TBM Data

As described in section 7.6.1, it would be very interesting to combine estimates of tensor based morphometry taken from subject to atlas, and longitudinal registrations. This may give a better estimate of the current or future disease state. This is because both AD and normal ageing are known to follow non-linear rates of atrophy [158]. Therefore, a more complex prediction model could be constructed that encompasses the current rate of change, as given by longitudinal imaging, and the current state, as estimated by the difference between a subject an atlas.

9.3 Final Conclusions

Formulating non-rigid registration within a fully Bayesian framework provides a powerful approach to medical image registration. The inference of the level of regularisation and data fidelity allows the flexible application of this approach to a variety of registration problems, with minimal manual intervention required. Describing registration through a probabilistic model allows intrinsic measures

of registration uncertainty. These uncertainty estimates have been demonstrated to provide information pertinent to improving spatially normalised statistical analysis. The use of fully Bayesian approaches to non-rigid registration is in its infancy, but further developments are likely due to the range of benefits it offers. Registration uncertainty in particular, is likely to attract wider interest as it forms an important consideration in any application where registration is required.

Appendix A

Variational Free Energy for the Probabilistic Registration Model

This appendix provides a derivation and definition of the negative variational free energy, $\mathcal{F}$, for the probabilistic registration model described in Chapter 3. As described in Section 2.2.4, $\mathcal{F}$ comprises two components:

$$\mathcal{F} = \mathcal{L}_{av} - D_{KL}(q(\mathbf{w}, \lambda, \phi) \| p(\mathbf{w}, \lambda, \phi)) \tag{A.1}$$

The mean-field approximation is used to separate the approximate posterior distribution, $q(\mathbf{w}, \phi, \lambda) = q(\mathbf{w})q(\lambda)q(\phi)$. $\mathcal{F}$ can now be written as:

$$\mathcal{F} = \mathcal{L}_{av} - D_{KL}(q(\mathbf{w}) \| p(\mathbf{w})) - D_{KL}(q(\lambda) \| P(\lambda)) - D_{KL}(q(\phi) \| P(\phi)) \tag{A.2}$$

where $q(\mathbf{w})$ is a multivariate normal distribution: $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Upsilon}^{-1})$. $q(\lambda)$ and $q(\phi)$ are Gamma distributed: $q(\lambda) = \mathrm{Ga}(\lambda; s, c)$ and $q(\phi) = \mathrm{Ga}(\phi; a, b)$.

The marginal log likelihood with respect to the model parameters, $\mathcal{L}_{av}$, is derived by integrating the log-likelihood, given in equation 3.4, over the approximate posterior distributions.

$$\begin{aligned} \mathcal{L}_{av} &= \int q(\mathbf{w})q(\lambda)q(\phi)(\log p(\mathbf{y}|\mathbf{w}, \lambda, \phi))d\mathbf{w}d\lambda d\phi \\ &= \langle \log p(\mathbf{y}|\mathbf{w}, \lambda, \phi) \rangle_{q(\mathbf{w})q(\lambda)q(\phi)} \\ &= \left\langle \frac{\alpha N_v}{2} \log \frac{\phi}{2\pi} - \frac{\alpha\phi}{2}(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w}))^{\mathsf{T}}(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})) \right\rangle_{q(\mathbf{w})q(\phi)} + \kappa \end{aligned} \tag{A.3}$$

where κ is a constant and α is the virtual decimation factor as described in Section 3.2.2. Consider the first term within the expectation bracket:

$$\left\langle \frac{\alpha N_v}{2} \log \frac{\phi}{2\pi} \right\rangle_{q(\phi)} = \frac{\alpha N_v}{2}(\log(a) + \psi(b)) - \frac{\alpha N_v}{2}(log 2\pi) \quad \text{(A.4)}$$

$$= \frac{\alpha N_v}{2}(\log(a) + \psi(b)) + \kappa \quad \text{(A.5)}$$

where ψ is the di-gamma function.

The second term is more complicated, as $\mathbf{w}$ cannot be directly integrated because it is a parameter in a non-linear function. Therefore, $\mathbf{t}(\mathbf{x}, \mathbf{w})$ needs to be approximated as linear. In this case, a 1st order Taylor series expansion is used:

$$\mathbf{t}(\mathbf{x}, \mathbf{w}) \approx \mathbf{t}(\mathbf{x}, \mu) + \mathbf{J}(\mathbf{w} - \mu) \quad \text{(A.6)}$$

where $\mathbf{J}$ is the Jacobian matrix of first order partial derivatives taken arround the current mean estimate. The entries of $\mathbf{J}$ can be calculated as:

$$\mathbf{J}_{i,j} = \frac{d(\mathbf{t}(\mathbf{x}, \mathbf{w})_i)}{d\mathbf{w}_j} \quad \text{(A.7)}$$

where $\mathbf{w}$ is taken as μ. i indexes voxels, and j indexes transformation parameters. As $\mathbf{t}(\mathbf{x}, \mathbf{w})$ only occurs in the context $\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})$, this is written as:

$$\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w}) \simeq \mathbf{y} - \mathbf{t}(\mathbf{x}, \mu) - \mathbf{J}(\mathbf{w} - \mu)$$

$$= \mathbf{k} - \mathbf{J}(\mathbf{w} - \mu) \quad \text{(A.8)}$$

where $\mathbf{k} = \mathbf{y} - \mathbf{t}(\mathbf{x}, \mu)$, and corresponds to a vectorisation of the residual image.

Finally, the second term of equation A.3 can be written as:

$$= \left\langle \frac{\alpha\phi}{2} \left(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})\right)^{\mathsf{T}} \left(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})\right) \right\rangle_{q(\mathbf{w})q(\phi)} \quad \text{(A.9)}$$

$$= \left\langle \frac{\alpha\phi}{2} (\mathbf{k} - \mathbf{J}(\mathbf{w} - \mu))^{\mathsf{T}} (\mathbf{k} - \mathbf{J}(\mathbf{w} - \mu)) \right\rangle_{q(\mathbf{w})q(\phi)}$$

$$= \left\langle \frac{\alpha\phi}{2} \right\rangle_{q(\phi)} \left(\left\langle \mathbf{k}^{\mathsf{T}}\mathbf{k} + (\mathbf{w} - \mu)^{\mathsf{T}} \mathbf{J}^{\mathsf{T}} \mathbf{J} (\mathbf{w} - \mu) \right\rangle_{q(\mathbf{w})} \right) + \kappa$$

$$= \frac{\alpha\bar{\phi}}{2} \left(\mathbf{k}^{\mathsf{T}}\mathbf{k} + \mathrm{Tr}(\Upsilon^{-1}\mathbf{J}^{\mathsf{T}}\mathbf{J}) \right) + \kappa \quad \text{(A.10)}$$

where $\bar{\phi}$ is the expectation of $q(\phi)$, $\bar{\phi} = ab$ and the integration result $\left\langle (\mathbf{a}-\mathbf{b})^{\mathsf{T}}\mathbf{C}(\mathbf{a}-\mathbf{b})\right\rangle_{\mathcal{N}(\mathbf{a};\mathbf{b},\mathbf{D})} = \mathrm{Tr}(\mathbf{DC}), \forall \mathbf{C}$ has been used. Tr is the matrix trace operator.

This leads to an expression for the marginal log-likelihood with respect to the model parameters:

$$\mathcal{L}_{av} = \frac{\alpha N_v}{2}(\log(a) + \psi(b)) - \frac{\alpha\bar{\phi}}{2}(\mathbf{k}^T\mathbf{k} + \mathrm{Tr}(\boldsymbol{\Upsilon}^{-1}\mathbf{J}^T\mathbf{J})) + \kappa \tag{A.11}$$

The Kullback-Leibler divergence terms D_{KL}, which penalises deviation of the approximate posterior from the prior distribution, are well established for Normal and Gamma distributions. The KL divergence between the $q(\phi)$ and $P(\phi)$ is given by:

$$\begin{aligned} D_{KL}(q(\phi)\|P(\phi)) &= (b-1)\psi(b) - \log(a) - b - \log(\Gamma(b)) + \log(\Gamma(b_0)) \\ &\quad + b_0\log a_0 - (b_0-1)(\psi(b) + \log(a)) + \frac{ab}{a_0} \end{aligned} \tag{A.12}$$

The KL divergence between $q(\lambda)$ and $P(\lambda)$is given by:

$$\begin{aligned} D_{KL}(q(\lambda)\|P(\lambda)) &= (c-1)\psi(c) - \log(s) - c - \log(\Gamma(c)) + \log(\Gamma(c_0)) \\ &\quad + c_0\log s_0 - (c_0-1)(\psi(c) + \log(s)) + \frac{sc}{s_0} \end{aligned} \tag{A.13}$$

The KL divergence between $q(\mathbf{w})$ and $p(\mathbf{w})$ is given by:

$$\begin{aligned} D_{KL}(q(\mathbf{w})\|p(\mathbf{w})) &= \frac{1}{2}\log\frac{|(\lambda\boldsymbol{\Lambda})^{-1}|}{|\boldsymbol{\Upsilon}^{-1}|} + \frac{\lambda}{2}Tr(\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Lambda}) + \frac{\lambda}{2}\mathbf{w}\boldsymbol{\Lambda}\mathbf{w} - \frac{N_c}{2} \\ &= \frac{1}{2}\left(-N_c\log\lambda - \log|\boldsymbol{\Lambda}| + \log|\boldsymbol{\Upsilon}|\right) + \frac{\lambda}{2}Tr(\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Lambda}) \\ &\quad + \frac{\lambda}{2}\mathbf{w}\boldsymbol{\Lambda}\mathbf{w} - \frac{N_c}{2} \end{aligned} \tag{A.14}$$

Appendix B

Derivation of Updates for Probabilistic Registration Model

This appendix presents the derivation of the updates for the probabilistic registration model given in Chapter 3. This derivation refers to the description of the VB calculus of variations procedure in Appendix 2.2.4 and optimises the definition of $\mathcal{F}$ given in Appendix A.

An appropriate form for the approximate posterior parameter distributions need to be selected. As stated in equation 2.29, an approximate posterior distribution, q, is optimised by making its logarithm equal to the sum of two terms.

As a local approximation of a linear relationship between the transformation parameters $\mathbf{w}$, and the Gaussian likelihood of the image residuals is used, the approximate posterior distribution on $\mathbf{w}$ is described as normally distributed, hence $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Upsilon}^{-1})$. The approximate log posterior for $\mathbf{w}$ is given as:

$$\log q(\mathbf{w}) \quad = \quad -\frac{1}{2}\mathbf{w}^T\boldsymbol{\Upsilon}\mathbf{w} + \frac{1}{2}\mathbf{w}^T\boldsymbol{\Upsilon}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Upsilon}\mathbf{w} + \kappa \tag{B.1}$$

where κ is used to describe all terms that are constant with respect to the parameter of interest, in this case, $\mathbf{w}$.

The posterior distribution on λ as a conjugate to the prior, is Gamma distributed, and parameterised using shape and scale parameters, c and s respectively. The approximate log posterior is given by:

$$\log q(\lambda) = -\frac{\lambda}{s} + (c-1)\log\lambda + \kappa \tag{B.2}$$

Similarly, the posterior distribution on ϕ is Gamma distributed, with shape and scale parameters b and a. The log posterior is given as:

$$\log q(\phi) = -\frac{\phi}{a} + (b-1)\log\phi + \kappa \tag{B.3}$$

Equation 2.29 provides a means for updating each separate approximate (log) posterior distribution to the local maximum of $\mathcal{F}$, for this model. In equation 2.29, it can be seen that the right hand side of the equality consists of two terms: The log-likehood of the data and the priors on the model parameters, both of which are taken with respect to the approximate posterior distributions on all the parameters except those that are being optimised. For the purposes of deriving the parameter updates, it is convenient to write the unmarginalised log-likelihood and priors as $\mathcal{M}$:

$$\begin{aligned}\mathcal{M} &= \log p(\mathbf{y}|\mathbf{x},\mathbf{w},\phi) + \log p(\mathbf{w}) + \log P(\lambda) + \log P(\phi) + \kappa \qquad \text{(B.4)}\\ &= \alpha\frac{N_v}{2}\log\phi - \frac{\phi\alpha}{2}(\mathbf{y}-\mathbf{t}(\mathbf{w}))^T(\mathbf{y}-\mathbf{t}(\mathbf{w})) - \frac{\lambda}{s_0} + (c_0-1)\log\lambda - \frac{\lambda}{2}\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}\\ &\quad + \frac{N_c}{2}\log\lambda - \frac{\phi}{a_0} + (b_0-1)\log\phi + \kappa\end{aligned}$$

where α is the virtual decimation factor as described in Section 3.2.2. $\mathcal{M}$ is used to derive updates for each parameter group, by marginalising over the other approximate posterior distributions and equating this with the log-posterior of the parameters of interest.

When using mean-field variational Bayesian inference with conjugate, exponential distributions, analytic updates can be derived for the hyper-parameters of each approximate posterior distribution by integrating the log posterior $\mathcal{M}$ over the other approximate posterior distributions. Updates are found by comparing the coefficients of the approximate log posterior hyper-parameters with those in the marginalised full log posterior.

Update on w

In order to derive the update for $\mathbf{w}$, the approximate posterior distributions of ϕ and λ needs to be integrated out from $\mathcal{M}$:

$$\log q(\mathbf{w}) = \int \mathcal{M} q(\lambda) q(\phi) d\lambda d\phi + \kappa \tag{B.5}$$

Substituting $\mathcal{M}$ and taking the expectation with respect to $q(\lambda)$ and $q(\phi)$:

$$\begin{aligned} &= \left\langle -\frac{\phi\alpha}{2}(\mathbf{y}-\mathbf{t}(\mathbf{x},\mathbf{w}))^T(\mathbf{y}-\mathbf{t}(\mathbf{x},\mathbf{w})) - \frac{\lambda}{2}\mathbf{w}^\mathsf{T}\mathbf{\Lambda}\mathbf{w} \right\rangle_{q(\lambda)q(\phi)} + \kappa \\ &= -\frac{1}{2}\{(\mathbf{y}-\mathbf{t}(\mathbf{x},\mathbf{w}))^\mathsf{T}(\mathbf{y}-\mathbf{t}(\mathbf{x},\mathbf{w}))\bar{\phi}\alpha + (\mathbf{w}^\mathsf{T}\mathbf{\Lambda}\mathbf{w})\bar{\lambda}\} + \kappa \end{aligned} \tag{B.6}$$

where the expectation of $q(\lambda)$ and $q(\phi)$ are used, and given as $\bar{\lambda}$ and $\bar{\phi}$. A linear approximation to the transformation $\mathbf{t}(\mathbf{x},\mathbf{w})$ is made using a first order Taylor series expansion as described in equation A.6. The right hand side of equation B.6 can be rewritten using the Taylor series expansion as:

$$\begin{aligned} &= -\frac{1}{2}\left\{\left[(\mathbf{k}-\mathbf{J}(\mathbf{w}-\boldsymbol{\mu}))^\mathsf{T}(\mathbf{k}-\mathbf{J}(\mathbf{w}-\boldsymbol{\mu}))\right]\bar{\phi}\alpha + \mathbf{w}^\mathsf{T}\mathbf{\Lambda}\mathbf{w}\bar{\lambda}\right\} + \kappa \\ &= \frac{1}{2}\left\{\bar{\phi}\alpha(\mathbf{J}^\mathsf{T}(\mathbf{k}+\mathbf{J}\boldsymbol{\mu})^\mathsf{T})\mathbf{w} + \bar{\phi}\alpha\mathbf{w}^\mathsf{T}(\mathbf{J}^\mathsf{T}(\mathbf{k}+\mathbf{J}\boldsymbol{\mu})) - (\mathbf{w}^\mathsf{T}(\mathbf{J}^\mathsf{T}\mathbf{J}\bar{\phi}\alpha + \mathbf{\Lambda}\bar{\lambda})\mathbf{w})\right\} + \kappa \end{aligned} \tag{B.7}$$

where $\mathbf{k}$ is the vectorisation of $\mathbf{y}-\mathbf{t}(\boldsymbol{\mu})$ and $\mathbf{J}$ is the Jacobian matrix of partial derivatives of each voxel in the image, with respect to each transformation parameter. By comparing the coefficients of $\mathbf{\Upsilon}$ and $\boldsymbol{\mu}$ between equation B.7 and the posterior distribution for $\mathbf{w}$ given in equation B.1, the hyper-parameter updates are given as:

$$\mathbf{\Upsilon} = \mathbf{J}^\mathsf{T}\mathbf{J}\bar{\phi}\alpha + \mathbf{\Lambda}\bar{\lambda} \tag{B.8}$$

$$\mathbf{\Upsilon}\boldsymbol{\mu}_{new} = \mathbf{J}^\mathsf{T}(\mathbf{J}\boldsymbol{\mu}_{old} + \mathbf{k})\bar{\phi}\alpha \tag{B.9}$$

where $\boldsymbol{\mu}_{new}$ and $\boldsymbol{\mu}_{old}$ are the new and old estimates for $\boldsymbol{\mu}$ respectively.

Updates for λ

To update λ, the posterior on $\mathbf{w}$ and ϕ is integrate out from the full log posterior:

$$\log q(\lambda) = \int \mathcal{M} q(\mathbf{w})q(\phi)d\mathbf{w}d\phi + \kappa \tag{B.10}$$

The right hand side of the above equation can be written as:

$$\begin{aligned} &= \left\langle -\frac{\lambda}{s_0} + \left(\frac{N_c}{2} + c_0 - 1\right)\log\lambda - \frac{\lambda}{2}\mathbf{w}^\mathsf{T}\mathbf{\Lambda}\mathbf{w} \right\rangle_{q(\mathbf{w})q(\phi)} + \kappa \\ &= -\frac{\lambda}{s_0} + (\frac{N_c}{2} + c_0 - 1)\log\lambda - \frac{\lambda}{2}\left(Tr(\mathbf{\Upsilon}^{-1}\mathbf{\Lambda}) + \boldsymbol{\mu}^\mathsf{T}\mathbf{\Lambda}\boldsymbol{\mu}\right) + \kappa \end{aligned} \quad \text{(B.11)}$$

where Tr refers to the matrix trace operation. By comparing coefficients of s and c from equation B.11 and the posterior distribution for $q(\lambda)$ given in equation B.2, the hyper-parameter updates are given as:

$$c = c_0 + \frac{N_c}{2} \quad \text{(B.12)}$$

$$\frac{1}{s} = \frac{1}{s_0} + \frac{1}{2}\left(\mathrm{Tr}(\mathbf{\Upsilon}^{-1}\mathbf{\Lambda}) + \boldsymbol{\mu}^\mathsf{T}\mathbf{\Lambda}\boldsymbol{\mu}\right) \quad \text{(B.13)}$$

Updates for ϕ

To update ϕ, posterior on $\mathbf{w}$ and λ can be integrated out from $\mathcal{M}$:

$$\log q(\phi) = \int \mathcal{M} q(\mathbf{w})q(\lambda)d\mathbf{w}d\lambda + \kappa \quad \text{(B.14)}$$

The right hand side of the above equation can be written as:

$$\begin{aligned} &= \left\langle \frac{\alpha N_v}{2}\log\phi - \frac{\phi}{a_0} - \frac{\phi\alpha}{2}(\mathbf{y} - \mathbf{t}(\mathbf{x},\mathbf{w}))^\mathsf{T}(\mathbf{y} - \mathbf{t}(\mathbf{x},\mathbf{w})) + (b_0 - 1)\log\phi \right\rangle_{q(\mathbf{w})q(\lambda)} + \kappa \\ &= \frac{N_v\alpha}{2}\log\phi + (b_0 - 1)\log\phi - \frac{\phi}{a_0} - \frac{\phi\alpha}{2}(\mathbf{k}^\mathsf{T}\mathbf{k} + \mathrm{Tr}(\mathbf{\Upsilon}^{-1}\mathbf{J}^\mathsf{T}\mathbf{J})) + \kappa \end{aligned} \quad \text{(B.15)}$$

By comparing coefficients of a and b from equation B.15 and the posterior distribution for $q(\phi)$ given in equation B.3, the hyper-parameter updates are given as:

$$b = b_0 + \frac{N_v\alpha}{2} \quad \text{(B.16)}$$

$$\frac{1}{a} = \frac{1}{a_0} + \frac{\alpha}{2}(\mathbf{k}^\mathsf{T}\mathbf{k} + \mathrm{Tr}(\mathbf{\Upsilon}^{-1}\mathbf{J}^\mathsf{T}\mathbf{J})) \quad \text{(B.17)}$$

Appendix C

Accuracy of approximate inference of λ and ϕ

As mentioned in Section 3.4.1, for the noise and spatial precision updates in equations 3.19 and 3.17, an approximation of control point independence is used. This is because the inverse of the large sparse precision matrix $\boldsymbol{\Upsilon}$ is required, and is computationally impractical to calculate or store. Therefore, this matrix is approximated as having independence between control points, only including the covariance between directions. The effect of this approximation is that when calculating the covariance matrix, the estimated variance will be lower in regions that contain a large amount of information, e.g. edges, and higher in homogeneous regions, which have a large amount of image covariance. This will induce a preference for slightly higher spatial precision, and lower noise precision. In order to test the accuracy of this approximation, a series of 200 2D registrations was carried out, resolving smooth artificial deformations of varying magnitudes applied to random image slices from the NIREP database. The results presented in figure C.1 show that this approximation produces reasonably accurate inference of these two parameters, with a 5-14% bias towards a higher spatial precision and a 3-7% bias to lower noise precision.

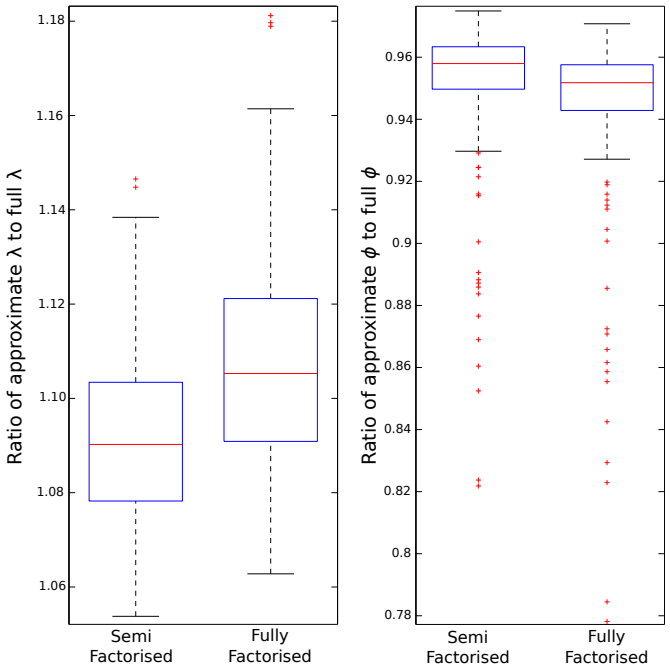

Figure C.1: Boxplots showing the accuracy of the approximation used in the inference of λ and ϕ. The boxplots show the ratio of the approximate inferred parameter to the inferred parameter using the full matrix inverse. The Semi Factorised plot refers to including the covariance between directions for each control point, whereas the Fully Factorised approach ignores all covariance. The registration experiments were performed in 2D, resolving random artificial deformations of a range of magnitudes. The plots show that λ is over-estimated by the semi factorised scheme by around 5-14% and ϕ is underestimated by 3-7%. It is also clear that the semi factorised scheme gives results that are closer to the truth, as would be expected.

Bibliography

[1] P. Aljabar, K.K. Bhatia, M. Murgasova, J.V. Hajnal, J.P. Boardman, L. Srinivasan, M.A. Rutherford, L.E. Dyet, A.D. Edwards, and D. Rueckert. Assessment of brain growth in early childhood using deformation-based morphometry. *NeuroImage*, 39(1):348 – 358, 2008.

[2] S. Allassonniére, Y. Amit, and A. Trouvè. Toward a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society, Series B*, 69(2), 2007.

[3] S. Allassonnière, E. Kuhn, and A. Trouvé. Construction of Bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.

[4] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, pages 376–387, 1991.

[5] J. Andersson, M. Jenkinson, and S Smith. Non-linear registration, aka spatial normalisation. *FMRIB technical report TR07JA1 from www.fmrib.ox.ac.uk/analysis/techrep*, 2007.

[6] N. Andreasen, L. Minthon, P. Davidsson, E. Vanmechelen, H. Vanderstichele, B. Winblad, and K. Blennow. Evaluation of CSF-tau and CSF-Aβ42 as diagnostic markers for Alzheimer disease in clinical practice. *Archives of Neurology*, 58(3):373, 2001.

[7] J. Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[8] J. Ashburner, J.L.R. Andersson, and K.J. Friston. High-dimensional image registration using symmetric priors. *NeuroImage*, 9(6):619–628, 1999.

[9] J. Ashburner, J.L.R. Andersson, and K.J. Friston. Image registration using a symmetric priorin three dimensions. *Human brain mapping*, 9(4):212–225, 2000.

[10] J. Ashburner and K.J. Friston. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–266, 1999.

[11] J. Ashburner and K.J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6):805–821, 2000.

[12] J. Ashburner and K.J. Friston. Why voxel-based morphometry should be used. *Neuroimage*, 14(6):1238–1243, 2001.

[13] J. Ashburner and K.J. Friston. Unified segmentation. *Neuroimage*, 26(3):839–851, 2005.

[14] J. Ashburner and K.J. Friston. Diffeomorphic registration using geodesic shooting and gauss-newton optimisation. *NeuroImage*, 55(3):954 – 967, 2011.

[15] J. Ashburner, P. Neelin, D. L. Collins, A. C. Evans, and K.J. Friston. Incorporating prior knowledge into image registration. *NeuroImage*, 6:344–352, 1997.

[16] H. Attias. A variational Bayesian framework for graphical models. *Advances in neural information processing systems*, 12(1-2):209–215, 2000.

[17] BB Avants, CL Epstein, M. Grossman, and JC Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[18] R. Bajcsy and S. Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1 – 21, 1989.

[19] R. Bajcsy, R. Lieberson, M. Reivich, et al. A computerized system for the elastic matching of deformed radiographic images to idealized atlas images. *Journal of computer assisted tomography*, 7(4):618, 1983.

[20] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415, 1975.

[21] M.F. Beg and A. Khan. Symmetric data attachment terms for large deformation image registration. *IEEE Transactions on Medical Imaging*, 26(9):1179–1189, 2007.

[22] M.F. Beg, M.I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005.

[23] KK Bhatia, J.V. Hajnal, BK Puri, AD Edwards, and D. Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *IEEE International Symposium on Biomedical Imaging*, pages 908–911. IEEE, 2004.

[24] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer, New York, 2006.

[25] D.J. Blezek and J.V. Miller. Atlas stratification. *Medical Image Analysis*, 11(5):443–457, 2007.

[26] F.L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567 –585, 1989.

[27] F.L. Bookstein. voxel-based morphometry should not be used with imperfectly registered images. *Neuroimage*, 14(6):1454–1462, 2001.

[28] A.P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[29] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[30] M. Bro-Nielsen and C. Gramkow. Fast fluid registration of medical images. In *Visualization in Biomedical Computing*, pages 265–276. Springer, 1996.

[31] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H.M. Arrighi. Forecasting the global burden of Alzheimers disease. *Alzheimer's and Dementia*, 3(3):186–191, 2007.

[32] O. Camara, J.A. Schnabel, G.R. Ridgway, W.R. Crum, A. Douiri, R.I. Scahill, D.L.G. Hill, and N.C. Fox. Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal Alzheimer's disease images. *NeuroImage*, 42(2):696–709, 2008.

[33] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[34] M.A. Chappell, A.R. Groves, B. Whitcher, and M.W. Woolrich. Variational Bayesian inference for a nonlinear forward model. *IEEE Transactions on Signal Processing*, 57(1):223 –236, 2009.

[35] G.E. Christensen, X. Geng, J.G. Kuhl, J. Bruss, T.J. Grabowski, I.A. Pirwani, M.W. Vannier, J.S. Allen, and H. Damasio. Introduction to the nonrigid image registration evaluation project (NIREP). *International Workshop on Biomedical Image Registration*, 4057:128–135, 2006.

[36] G.E. Christensen and H.J. Johnson. Consistent image registration. *IEEE Transactions on Medical Imaging*, 20(7):568–582, 2001.

[37] G.E. Christensen, R.D. Rabbitt, and M.I. Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5(10):1435–1447, 1996.

[38] M.K. Chung, K.J. Worsley, T. Paus, C. Cherif, D.L. Collins, J.N. Giedd, J.L. Rapoport, and A.C. Evans. A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14(3):595 – 606, 2001.

[39] D.L. Collins, P. Neelin, T.M. Peters, A.C. Evans, et al. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of computer assisted tomography*, 18(2):192, 1994.

[40] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[41] R.T. Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13, 1946.

[42] WR Crum, LD Griffin, DLG Hill, and DJ Hawkes. Zen and the art of medical image registration: correspondence, homology, and quality. *NeuroImage*, 20(3):1425–1437, 2003.

[43] W.R. Crum, R.I. Scahill, and N.C. Fox. Automated hippocampal segmentation by regional fluid registration of serial MRI: validation and application in Alzheimer's disease. *Neuroimage*, 13(5):847–855, 2001.

[44] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.O. Habert, M. Chupin, H. Benali, and O. Colliot. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *Neuroimage*, 56(2):766–781, 2011.

[45] C. Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Computer Vision and Image Understanding*, 66(2):207–222, 1997.

[46] G. De Meyer, F. Shapiro, H. Vanderstichele, E. Vanmechelen, S. Engelborghs, P.P. De Deyn, E. Coart, O. Hansson, L. Minthon, H. Zetterberg, et al. Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Archives of Neurology*, 67(8):949, 2010.

[47] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[48] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[49] T. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15, 2000.

[50] B. Dubois, H.H. Feldman, C. Jacova, J.L. Cummings, S.T. DeKosky, P. Barberger-Gateau, A. Delacourte, G. Frisoni, N.C. Fox, D. Galasko, et al. Revising the definition of Alzheimer's disease: a new lexicon. *The Lancet Neurology*, 9(11):1118–1127, 2010.

[51] J. Dukart, M.L. Schroeter, and K. Mueller. Age correction in dementia–matching to a healthy brain. *PloS ONE*, 6(7):e22193, 2011.

[52] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1):1–26, 1979.

[53] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

[54] A.C. Evans, D.L. Collins, SR Mills, ED Brown, RL Kelly, and TM Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, pages 1813–1817. IEEE, 1993.

[55] B. Fischl, M.I. Sereno, R.B.H. Tootell, A.M. Dale, et al. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4):272–284, 1999.

[56] N.C. Fox, R.S. Black, S. Gilman, M.N. Rossor, S.G. Griffith, L. Jenkins, and M. Koller. Effects of Aβ immunization (AN1792) on MRI measures of cerebral volume in Alzheimer disease. *Neurology*, 64(9):1563–1572, 2005.

[57] N.C. Fox and J.M. Schott. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet*, 363:392394, 2004.

[58] Peter T. Fox. Spatial normalization origins: Objectives, applications, and alternatives. *Human Brain Mapping*, 3(3):161–164, 1995.

[59] P.A. Freeborough and N.C. Fox. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Transactions on Medical Imaging*, 16(5):623–629, 1997.

[60] P.A. Freeborough and N.C. Fox. Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images. *Journal of Computer Assisted Tomography*, 22(5):838, 1998.

[61] Giovanni B Frisoni, Nick C Fox, Clifford R Jack, Philip Scheltens, and Paul M Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature reviews. Neurology*, 6(2):67–77, 2010.

[62] K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattout. Multiple sparse priors for the M/EEG inverse problem. *NeuroImage*, 39(3):1104–1120, 2008.

[63] Karl. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J. Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189, 1995.

[64] J.C. Gee and R.K. Bajcsy. Elastic matching: Continuum mechanical and probabilistic analysis. *Brain warping*, 183:197, 1999.

[65] J.C. Gee, D.R. Haynor, M. Reivich, and R. Bajcsy. Finite element approach to warping of brain images. In M.H. Loew, editor, *Medical Imaging 1994*, volume 2167. SPIE, 1994.

[66] J.C. Gee, L. Le Briquer, C. Barillot, and DR Haynor. Probabilistic matching of brain images. In *Information processing in medical imaging*, volume 3, pages 113–125, 1995.

[67] J.C. Gee, L. LeBriquer, C. Barillot, D.R. Haynor, and R. Bajcsy. Bayesian approach to the brain image matching problem. *IRCS Technical Reports Series*, page 123, 1995.

[68] J.C. Gee, M. Reivich, and R. Bajcsy. Elastically deforming a three-dimensional atlas to match anatomical brain images. *Journal of computer assisted tomography*, 17:225–236, 1993.

[69] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian analysis*, 1(3):515–533, 2006.

[70] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[71] Catriona D. Good, Ingrid S. Johnsrude, John Ashburner, Richard N.A. Henson, Karl J. Friston, and Richard S.J. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14(1):21 – 36, 2001.

[72] Adrian R. Groves, Michael A. Chappell, and Mark W. Woolrich. Combined spatial and non-spatial prior for inference on MRI time-series. *NeuroImage*, 45(3):795 – 809, 2009.

[73] A.R. Groves, C.F. Beckmann, S.M. Smith, and M.W. Woolrich. Linked independent component analysis for multimodal data fusion. *NeuroImage*, 54(3):2198 – 2217, 2011.

[74] H. Gudbjartsson and S. Patz. The Rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6):910–914, 1995.

[75] A. Guimond, J. Meunier, and J.P. Thirion. Average brain models: A convergence study. *Computer vision and image understanding*, 77(2):192–210, 2000.

[76] H. Hampel, K. Burger, J.C. Pruessner, R. Zinkowski, J. DeBernardis, D. Kerkman, G. Leinsinger, A.C. Evans, P. Davies, H.J. Moller, et al. Correlation of cerebrospinal fluid levels of tau protein phosphorylated at threonine 231 with rates of hippocampal atrophy in Alzheimer disease. *Archives of neurology*, 62(5):770, 2005.

[77] J.A. Hanley, B.J. McNeil, et al. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[78] LM Harrison, W. Penny, J. Daunizeau, and KJ Friston. Diffusion-based spatial priors for functional magnetic resonance images. *Neuroimage*, 41(2):408–423, 2008.

[79] L.M. Harrison, W. Penny, G. Flandin, C.C. Ruff, N. Weiskopf, and K.J. Friston. Graph-partitioned spatial priors for functional magnetic resonance images. *NeuroImage*, 43(4):694 – 707, 2008.

[80] R.H. Hashemi, W.G. Bradley, and C.J. Lisanti. *MRI: the basics*. Lippincott Williams & Wilkins, 2010.

[81] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[82] K. Herholz, E. Salmon, D. Perani, JC Baron, V. Holthoff, L. Frölich, P. Schönknecht, K. Ito, R. Mielke, E. Kalbe, et al. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. *Neuroimage*, 17(1):302–316, 2002.

[83] G. Hermosillo, C. Chefd'Hotel, and O. Faugeras. Variational methods for multimodal image matching. *International Journal of Computer Vision*, 50(3):329–343, 2002.

[84] D.L.G. Hill, P.G. Batchelor, M. Holden, and D.J. Hawkes. Medical image registration. *Physics in medicine and biology*, 46, 2001.

[85] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[86] M. Holden. A review of geometric transformations for nonrigid body registration. *IEEE Transactions on Medical Imaging*, 27(1):111–128, 2008.

[87] C.W. Hsu, C.C. Chang, and C.J. Lin. A practical guide to support vector classification, 2003.

[88] X. Hua, S. Lee, D.P. Hibar, I. Yanovsky, A.D. Leow, A.W. Toga, C.R. Jack Jr, M.A. Bernstein, E.M. Reiman, D.J. Harvey, et al. Mapping Alzheimer's disease progression in 1309 MRI scans: power estimates for different inter-scan intervals. *Neuroimage*, 51(1):63–75, 2010.

[89] X. Hua, S. Lee, I. Yanovsky, A.D. Leow, Y.Y. Chou, A.J. Ho, B. Gutman, A.W. Toga, C.R. Jack Jr, M.A. Bernstein, et al. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: an ADNI study of 515 subjects. *Neuroimage*, 48(4):668–681, 2009.

[90] X. Hua, A.D. Leow, N. Parikshak, S. Lee, M.C. Chiang, A.W. Toga, C.R. Jack Jr, M.W. Weiner, P.M. Thompson, et al. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. *Neuroimage*, 43(3):458–469, 2008.

[91] Xue Hua, Boris Gutman, Christina P. Boyle, Priya Rajagopalan, Alex D. Leow, Igor Yanovsky, Anand R. Kumar, Arthur W. Toga, Clifford R. Jack Jr., Norbert Schuff, Gene E. Alexander, Kewei Chen, Eric M. Reiman, Michael W. Weiner, and Paul M. Thompson. Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. *NeuroImage*, 57(1):5 – 14, 2011.

[92] X. Huang, Y. Sun, D. Metaxas, F. Sauer, and C. Xu. Hybrid image registration based on configural matching of scale-invariant salient region features. In *IEEE Cpmferemce on Computer Vision and Pattern Recognition*, pages 167–167. IEEE, 2004.

[93] M. Hub, M.L. Kessler, and C.P. Karger. A stochastic approach to estimate the uncertainty involved in B-spline image registration. *IEEE Transactions on Medical Imaging*, 28(11):1708 –1716, 2009.

[94] J.E. Iglesias, M.R. Sabuncu, and K. Van Leemput. Incorporating parameter uncertainty in Bayesian segmentation models: application to hippocampal subfield volumetry. In Nicholas Ayache, Herv Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention*, volume 7512 of *LNCS*, pages 50–57. Springer, Heidelberg, 2012.

[95] T.S. Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, pages 129–160, 2001.

[96] C.R. Jack, M. Slomkowski, S. Gracon, TM Hoover, J.P. Felmlee, K. Stewart, Y. Xu, M. Shiung, PC OBrien, R. Cha, et al. MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. *Neurology*, 60(2):253–260, 2003.

[97] C.R. Jack Jr, R.C. Petersen, Y.C. Xu, S.C. Waring, P.C. O'Brien, E.G. Tangalos, G.E. Smith, R.J. Ivnik, and E. Kokmen. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, 49(3):786–794, 1997.

[98] M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

[99] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[100] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.

[101] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[102] B Karacali and Christos Davatzikos. Estimating topology preserving and smooth displacement fields. *IEEE Transactions on Medical Imaging*, 23(7):868–880, 2004.

[103] M. Keller, A. Roche, A. Tucholka, and B. Thirion. Dealing with spatial normalization errors in fMRI group inference using hierarchical modeling. *Statistica Sinica*, 18(4):1357–1374, 2008.

[104] M. Kim, G. Wu, P.T. Yap, and D. Shen. A generalized learning based framework for fast brain image registration. *Medical Image Computing and Computer-Assisted Intervention*, pages 306–314, 2010.

[105] A. Klein, J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, M.C. Chiang, G.E. Christensen, D.L. Collins, J. Gee, P. Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786–802, 2009.

[106] S. Klöppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jack Jr, J. Ashburner, and R.S.J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008.

[107] W.E. Klunk, H. Engler, A. Nordberg, Y. Wang, G. Blomqvist, D.P. Holt, M. Bergström, I. Savitcheva, G.F. Huang, S. Estrada, et al. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Annals of neurology*, 55(3):306–319, 2004.

[108] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[109] S. Lee, G. Wolberg, and S.Y. Shin. Scattered data interpolation with multilevel B-splines. *IEEE Transactions on Visualization and Computer Graphics*, 3(3):228–244, 1997.

[110] A. Leow, S.C. Huang, A. Geng, J. Becker, S. Davis, A. Toga, and P. Thompson. Inverse consistent mapping in 3d deformable image registration: its construction and statistical properties. In *Information Processing in Medical Imaging*, pages 23–57. Springer, 2005.

[111] A.D. Leow, I. Yanovsky, M.C. Chiang, A.D. Lee, A.D. Klunder, A. Lu, J.T. Becker, S.W. Davis, A.W. Toga, and P.M. Thompson. Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration. *IEEE Transactions on Medical Imaging*, 26(6):822–832, 2007.

[112] A.D. Leow, I. Yanovsky, N. Parikshak, X. Hua, S. Lee, A.W. Toga, C.R. Jack Jr, M.A. Bernstein, P.J. Britson, J.L. Gunter, et al. Alzheimer's disease neuroimaging initiative: a one-year follow up study using tensor-based morphometry correlating degenerative rates, biomarkers and cognition. *Neuroimage*, 45(3):645–655, 2009.

[113] N. Lepore, C. Brun, Y.Y. Chou, M.C. Chiang, R.A. Dutton, K.M. Hayashi, E. Luders, O.L. Lopez, H.J. Aizenstein, A.W. Toga, et al. Generalized tensor-based morphometry of hiv/aids using multivariate statistics on deformation tensors. *IEEE Transactions on Medical Imaging*, 27(1):129–141, 2008.

[114] H. Lester, S. Arridge, K. Jansons, L. Lemieux, J. Hajnal, and A. Oatridge. Non-linear registration with the variable viscosity fluid algorithm. In Attila Kuba, Martin amal, and Andrew Todd-Pokropek, editors, *Information Processing in Medical Imaging*, volume 1613, pages 238–251. 1999.

[115] H. Lester and S.R. Arridge. A survey of hierarchical non-linear medical image registration. *Pattern recognition*, 32(1):129–149, 1999.

[116] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens. Nonrigid image registration using conditional mutual information. *IEEE Transactions on Medical Imaging*, 29(1):19–29, 2010.

[117] H. Luan, F. Qi, Z. Xue, L. Chen, and D. Shen. Multimodality image registration by maximization of quantitative-qualitative measure of mutual information. *Pattern Recognition*, 41(1):285–298, 2008.

[118] Grundman M, Petersen RC, Ferris SH, and et al. Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials. *Archives of Neurology*, 61(1):59–66, 2004.

[119] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

[120] Benot Magnin, Lilia Mesrob, Serge Kinkingnhun, Mlanie Plgrini-Issac, Olivier Colliot, Marie Sarazin, Bruno Dubois, Stphane Lehricy, and Habib Benali. Support vector machine-based classification of Alzheimers disease from whole-brain anatomical MRI. *Neuroradiology*, 51:73–83, 2009.

[121] J.B.Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1 – 36, 1998.

[122] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[123] C.E. Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[124] M. Miller, A. Banerjee, G. Christensen, S. Joshi, N. Khaneja, U. Grenander, and L. Matejic. Statistical methods in computational anatomy. *Statistical Methods in Medical Research*, 6(3):267, 1997.

[125] M.I. Miller, G.E. Christensen, Y. Amit, and U. Grenander. Mathematical textbook of deformable neuroanatomies. *Proceedings of the National Academy of Sciences*, 90(24):11944, 1993.

[126] M. Modat, G.. Ridgway, P. Daga, M. J. Cardoso, D. J. Hawkes, and S. Ashburner, J.and Ourselin. Log-euclidean free-form deformation. *Proceedings of SPIE*, pages 79621Q–79621Q–6, 2011.

[127] J. Modersitzki. *FAIR: flexible algorithms for image registration*, volume 6. Society for Industrial and Applied Mathematics (SIAM), 2009.

[128] J.C. Morris. The clinical dementia rating (cdr): current version and scoring rules. *Neurology; Neurology*, 1993.

[129] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C.R. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, and L. Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's and Dementia*, 1(1):55–66, 2005.

[130] H.H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):565–593, 1986.

[131] T. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446, 2003.

[132] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical Image Analysis*, 15(4):622 – 639, 2011.

[133] HD Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545, 1971.

[134] C.A. Pelizzari, GT Chen, D.R. Spelbring, R.R. Weichselbaum, and C.T. Chen. Accurate three-dimensional registration of CT, PET, and/or MR images of the brain. *Journal of computer assisted tomography*, 13(1):20, 1989.

[135] W.D. Penny, N.J. Trujillo-Barreto, and K.J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

[136] R.C. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004.

[137] C. Peterson and J.R. Anderson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1:995–1019, 1987.

[138] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. Numerical recipes in c: the art of scientific programming. *Section*, 15:682–683, 1992.

[139] M. Reuter, H.D. Rosas, and B. Fischl. Highly accurate inverse consistent registration: a robust approach. *Neuroimage*, 53(4):1181–1196, 2010.

[140] Martin Reuter and Bruce Fischl. Avoiding asymmetry-induced bias in longitudinal image processing. *NeuroImage*, 57(1):19 – 21, 2011.

[141] G.R. Ridgway. Statistical analysis for longitudinal MR imaging of dementia. *PhD Thesis, University College London*, 2009.

[142] B.H. Ridha, J. Barnes, J.W. Bartlett, A. Godbolt, T. Pepple, M.N. Rossor, and N.C. Fox. Tracking atrophy progression in familial Alzheimer's disease: a serial MRI study. *Lancet Neurology*, 5:828834, 2006.

[143] P. Risholm, J. Balter, and W. Wells. Estimation of delivered dose in radiotherapy: the influence of registration uncertainty. *Medical Image Computing and Computer-Assisted Intervention*, pages 548–555, 2011.

[144] P. Risholm, J. Balter, and W. Wells. Estimation of delivered dose in radiotherapy: the influence of registration uncertainty. In G. Fichtinger, A. Martel, and T. Peters, editors, *Medical Image Computing and Computer-Assisted Intervention*, LNCS, pages 548–555. Springer, Heidelberg, 2011.

[145] P. Risholm, A. Fedorov, J. Pursley, K. Tuncali, R. Cormack, and W.M. Wells. Probabilistic non-rigid registration of prostate images: Modeling and quantifying uncertainty. In *IEEE International Symposium on Biomedical Imaging*, pages 553 –556, 2011.

[146] P. Risholm, E. Samset, and W. Wells. Bayesian estimation of Ddformation and elastic parameters in non-rigid registration. *Workshop on Biomedical Image Registration*, pages 104–115, 2010.

[147] Petter Risholm, Steve Pieper, Eigil Samset, and William Wells. Summarizing and visualizing uncertainty in non-rigid registration. In Tianzi Jiang, Nassir Navab, Josien Pluim, and Max Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention 2010*, volume 6362, pages 554–561. 2010.

[148] A. Roche and N. Malandain, G.and Ayache. Unifying maximum likelihood approaches in medical image registration. *International Journal of Imaging Systems and Technology*, 11(1):71–80, 2000.

[149] Alexis Roche, Grégoire Malandain, Xavier Pennec, and Nicholas Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In *MICCAI*, pages 1115–1124. Springer, 1998.

[150] P. Rogelj and S. Kovacic. Symmetric image registration. *Medical Image Analysis*, 10(3):484–493, 2006.

[151] T. Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *IEEE Transactions on Medical Imaging*, 31(2):153 –163, feb. 2012.

[152] T. Rohlfing, D.B. Russakoff, and C.R. Maurer. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Transactions on Medical Imaging*, 23(8):983–994, 2004.

[153] S.A.R.B. Rombouts, F. Barkhof, R. Goekoop, C.J. Stam, and P. Scheltens. Altered resting state networks in mild cognitive impairment and mild Alzheimer's disease: an fMRI study. *Human brain mapping*, 26(4):231–239, 2005.

[154] S.E. Rose, F. Chen, J.B. Chalk, F.O. Zelaya, W.E. Strugnell, M. Benson, J. Semple, and D.M. Doddrell. Loss of connectivity in Alzheimer's disease: an evaluation of white matter tract integrity with colour coded MR diffusion tensor imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 69(4):528–530, 2000.

[155] A. Rosenfeld and A.C. Kak. Digital picture processing. *New York, Academic Press, 1982,*, 1982.

[156] D. Rueckert, LI Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using Free-Form Deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[157] M.R. Sabuncu, S.K. Balci, M.E. Shenton, and P. Golland. Image-driven population analysis through mixture modeling. *IEEE Transactions on Medical Imaging*, 28(9):1473–1487, 2009.

[158] RI Scahill and NC Fox. Longitudinal imaging in dementia. *British Journal of Radiology*, 80(Special Issue 2):S92–S98, 2007.

[159] R.I. Scahill, J.M. Schott, J.M. Stevens, M.N. Rossor, and N.C. Fox. Mapping the evolution of regional atrophy in Alzheimer's disease: unbiased analysis of fluid-registered serial MRI. *Proceedings of the National Academy of Sciences*, 99(7):4703, 2002.

[160] J. Schnabel, D. Rueckert, M. Quist, J. Blackall, A. Castellano-Smith, T. Hartkens, G. Penney, W. Hall, H. Liu, and C. Truwit. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. *Medical Image Computing and Computer-Assisted Intervention*, pages 573–581, 2001.

[161] J.A. Schnabel, C. Tanner, A.D. Castellano-Smith, A. Degenhard, M.O. Leach, D.R. Hose, D.L.G. Hill, and D.J. Hawkes. Validation of nonrigid image registration using finite-element methods: application to breast MR images. *IEEE Transactions on Medical Imaging*, 22(2):238–247, 2003.

[162] D. Shen and C. Davatzikos. HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, 21(11):1421–1439, 2002.

[163] I. J. A. Simpson, J. A. Schnabel, A. R. Groves, J. Andersson, and M. W. Woolrich. A Bayesian approach to non-rigid registration with adaptive regularisation. *Proceedings of the Organisation for Human Brain Mapping*, 2010.

[164] I. J. A Simpson, M.W. Woolrich, A.R. Groves, and J.A. Schnabel. Longitudinal Brain MRI Analysis with Uncertain Registration. In G. Fichtinger, A. Martel, and T. Peters, editors, *Medical Image Computing and Computer-Assisted Intervention 2011*, LNCS. Springer, Heidelberg, 2011.

[165] I. J. A Simpson, M.W. Woolrich, A.R. Groves, and J.A. Schnabel. Probabilistic segmentation propagation from uncertainty in registration. *Proceedings of Medical Image Understanding and Analysis*, 2011.

[166] I.J.A Simpson, J.A. Schnabel, J.L.R. Andersson, A.R. Groves, and M.W. Woolrich. Ensemble learning incorporating uncertain registration. *Proceeding of MIUA*, 2012.

[167] I.J.A. Simpson, J.A. Schnabel, J.L.R. Andersson, A.R. Groves, and M.W. Woolrich. A probabilistic non-rigid registration framework using local noise estimates. *Proceedings of IEEE International Symposium on Biomedical Imaging 2012*, pages 688–691, 2012.

[168] I.J.A Simpson, J.A. Schnabel, J.L.R. Andersson, A.R. Groves, and M.W. Woolrich. Ensemble learning incorporating uncertain registration. *IEEE Transaction on Medical Imaging*, 32:748 – 756, 2013.

[169] I.J.A. Simpson, J.A. Schnabel, A.R. Groves, J.L.R. Andersson, and M.W. Woolrich. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 59:2438–51, 2012.

[170] S.M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[171] S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, TE Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208, 2004.

[172] S.M. Smith, Y. Zhang, M. Jenkinson, J. Chen, PM Matthews, A. Federico, and N. De Stefano. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage*, 17(1):479–489, 2002.

[173] M. Staring, S. Klein, and J.P.W. Pluim. Nonrigid registration with tissue-dependent filtering of the deformation field. *Physics in medicine and biology*, 52:6879, 2007.

[174] C. Studholme and V. Cardenas. A template free approach to volumetric spatial normalization of brain anatomy. *Pattern Recognition Letters*, 25(10):1191–1202, 2004.

[175] J.A. Swets and R.M. Pickett. *Evaluation of diagnostic systems: Methods from signal detection theory*. Academic Press New York, 1982.

[176] Hemant Tagare, David Groisser, and Oskar Skrinjar. Symmetric non-rigid registration: A geometric theory and some numerical techniques. *Journal of Mathematical Imaging and Vision*, 34:61–88, 2009. 10.1007/s10851-008-0129-7.

[177] J. Talairach and P. Tournoux. *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging*. Thieme New York, 1988.

[178] L. Tang, G. Hamarneh, and R Abugharbieh. Reliability-driven, spatially-adaptive regularization for deformable registration. *Workshop on Biomedical Image Registration*, pages 173–185, 2010.

[179] S.J. Teipel, C. Born, M. Ewers, A.L.W. Bokde, M.F. Reiser, H.J. Möller, and H. Hampel. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *Neuroimage*, 38(1):13–24, 2007.

[180] J.P. Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[181] Wesley K. Thompson and Dominic Holland. Bias in tensor based morphometry stat-roi measures may result in unrealistic power estimates. *NeuroImage*, 57(1):1 – 4, 2011.

[182] M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

[183] K. Van Leemput. Encoding probabilistic brain atlases using Bayesian inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837, 2009.

[184] P. Vemuri, J.L. Gunter, M.L. Senjem, J.L. Whitwell, K. Kantarci, D.S. Knopman, B.F. Boeve, R.C. Petersen, and C.R. Jack Jr. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage*, 39(3):1186–1197, 2008.

[185] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Symmetric log-domain diffeomorphic registration: A demons-based approach. *Medical Image Computing and Computer-Assisted Intervention*, pages 754–761, 2008.

[186] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1, Supplement 1):S61 – S72, 2009.

[187] P. Viola and W.M. Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[188] U. Vovk, F. Pernus, and B. Likar. A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 26(3):405–421, 2007.

[189] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–21, July 2004.

[190] B.L. Welch. The generalization of Student's' problem when several different population variances are involved. *Biometrika*, pages 28–35, 1947.

[191] A.P. Witkin. Scale-space filtering. *Readings in computer vision: issues, problems, principles, and paradigms*, pages 329–332, 1987.

[192] R.P. Woods, S.R. Cherry, J.C. Mazziotta, et al. Rapid automated algorithm for aligning and reslicing PET images. *Journal of Computer Assisted Tomography*, 16(4):620, 1992.

[193] R.P. Woods, S.T. Grafton, C.J. Holmes, S.R. Cherry, and J.C. Mazziotta. Automated image registration: I. General methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography*, 22(1):139, 1998.

[194] M. Woolrich. Robust group analysis using outlier inference. *Neuroimage*, 41(2):286–301, 2008.

[195] MW Woolrich, TE Behrens, and O. FMRIB. Variational Bayes inference of spatial mixture models for segmentation. *IEEE transactions on medical imaging*, 25(10):1380–1391, 2006.

[196] M.W. Woolrich, M. Jenkinson, J.M. Brady, and S.M. Smith. Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE transactions on medical imaging*, 23(2):213–231, 2004.

[197] K.J. Worsley, S. Marrett, P. Neelin, and AC Evans. Searching scale space for activation in PET images. *Human brain mapping*, 4(1):74–90, 1996.

[198] K.J. Worsley, J.B. Poline, AC Vandal, and KJ Friston. Tests for distributed, nonfocal brain activations. *NeuroImage*, 2(3):183–194, 1995.

[199] IC Wright, PK McGuire, J.B. Poline, JM Travere, RM Murray, CD Frith, RSJ Frackowiak, and KJ Friston. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *Neuroimage*, 2(4):244–252, 1995.

[200] I. Yanovsky, A.D. Leow, S. Lee, S.J. Osher, and P.M. Thompson. Comparing registration methods for mapping brain change using tensor-based morphometry. *Medical image analysis*, 13(5):679–700, 2009.

[201] B.T.T. Yeo, M.R. Sabuncu, R. Desikan, B. Fischl, and P. Golland. Effects of registration regularization and atlas sharpness on segmentation accuracy. *Medical image analysis*, 12(5):603, 2008.

[202] B.T.T. Yeo, M.R. Sabuncu, T. Vercauteren, D.J. Holt, K. Amunts, K. Zilles, P. Golland, and B. Fischl. Learning task-optimal registration cost functions for localizing cytoarchitecture and function in the cerebral cortex. *IEEE Transactions on Medical Imaging*, 29(7):1424–1441, 2010.

[203] P.A. Yushkevich, B.B. Avants, M. S.R. Das amd Pluta, J. amd Altinay, and C. Craige. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3T MRI data. *Neuroimage*, 50:434445, 2010.

[204] H. Zhang. The optimality of naive bayes. *American Association for Artificial Intelligence*, 1(2):3, 2004.

[205] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[206] D. Zikic, M. Baust, A. Kamen, and N. Navab. Generalization of deformable registration in riemannian sobolev spaces. *Medical Image Computing and Computer-Assisted Intervention*, pages 586–593, 2010.

[207] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000, 2003.

[208] L. Zöllei, M. Jenkinson, S. Timoner, and W. Wells. A marginalized MAP approach and EM optimization for pair-wise registration. In *Information Processing in Medical Imaging*, pages 662–674. Springer, 2007.

[209] L. Zöllei, E. Learned-Miller, E. Grimson, and W. Wells. Efficient population registration of 3d data. *Computer Vision for Biomedical Image Applications*, pages 291–301, 2005.

[210] K.H. Zou, W.M. Wells III, R. Kikinis, and S.K. Warfield. Three validation metrics for automated probabilistic image segmentation of brain tumours. *Statistics in medicine*, 23(8):1259–1282, 2004.